EUR 4256 e

EUROPEAN ATOMIC ENERGY COMMUNITY - EURATOM

1968 MEETING OF EUROPEAN LIBRARIANS WORKING IN THE NUCLEAR FIELD

A selection of papers read at the 5th annual meeting of scientific librarians organized by the Centre for Information and Documentation (CID) and the Scientific Information Processing Center (CETIS) at Stresa (Italy), April 24-25, 1968

1969

Directorate general Dissemination of Information
Centre for Information and Documentation - CID
Luxemburg

LEGAL NOTICE

This report is on sale at the addresses listed on cover page 4

at the price of	FF 16,50	FB 165	DM 13.20	Lit. 2060	Fl. 11.95

When ordering, please quote the EUR number and the title, which are indicated on the cover of each report.

Printed by Vanmelle
Brussels, June 1969

CORRIGENDUM

EUR 4256 e

1968 Meeting of European librarians working in the nuclear field.

1. This meeting has been held in ISPRA and not in Stresa, as mentioned on the cover, title page and abstract cards.

2. The report has been published by :
Directorate General Dissemination of Information (CID - Luxemburg)
and
Directorate General Joint Research Center (CETIS - Ispra)

3. In the contribution of C. Vermeb, please add after page 100, the annexed page 100 bis.

7) For each item the computer adds up the weights of the corresponding index terms, the sum establishing the value of the item.

8) The computer prints the reference numbers, ranked by decreasing values.

9) The printout is cut-off at a number of references which is calculated as the product of the total number of answers to the "medium" query and the relevance ratio determined from the screened sample.

10) All references, down to the cut-off, are sent to the customer without any further screening.

Figure 15 presents two examples for which all the calculations and sorting operations to be done by the computer were effected manually.

The first example deals with "Uranium Isotope Separation". There were 220 answers to a "medium" query, 41 of which were relevant to the query (relevance ratio : 18.6 %). The points on the first line in Fig. 15 represent these "hits". (The references are arranged according to their sequence number). - The first 32 references were checked for relevancy ; 6 were found to be relevant (relevance ratio : 18.7 %. This corresponds very well with the overall relevancy of 18.6 %). The 6 hits and the 26 non-relevant references were used as a basis for the above described weighting and ranking procedure. The result of this procedure is shown in the second line. The hits were shifted to the left side, that means they are found preferably under the first references. The cut at 18.7 x 220 = 41 references yielded 32 hits (relevance ratio : 78 %). With three "clear" hits and three "clear" non-relevant references selected from the sample of 32 references the result was even a little bit better.

The second example refers to "Portable Radiation Detectors". There were 114 answers on a "medium" query, 18 of which were relevant to the query (relevance ratio : 15.8). The points on the third line in Fig. 15 represent these hits. The last 20 references were checked for relevancy : 3 were found to be relevant (relevance ratio : 15 %). This again corresponds very well to the overall relevance ...

EUR 4256 e

EUROPEAN ATOMIC ENERGY COMMUNITY - EURATOM

1968 MEETING OF EUROPEAN LIBRARIANS WORKING IN THE NUCLEAR FIELD

A selection of papers read at the 5th annual meeting of scientific librarians organized by the Centre for Information and Documentation (CID) and the Scientific Information Processing Center (CETIS) at Stresa (Italy), April 24-25, 1968

1969

Directorate general Dissemination of Information
Centre for Information and Documentation - CID
Luxemburg

VORWORT

Jahrestreffen europäischer, wissenschaftlicher Bibliothekare und Dokumentalisten die auf dem Nukleargebiet arbeiten, ist zur Tradition, ja fast zu einer feststehenden Einrichtung geworden, die den grossen Vorteil bietet, dass auf diese Weise jedes Jahr den Verantwortlichen der wissenschaftlichen Bibliotheken Europas die Möglichkeit gegeben wird, Informationen über die neuesten Entwicklungen auf den sie interessierenden Gebieten auszutauschen und über ihre Pläne für die nähere und fernere Zukunft zu sprechen.

Der Grund für die diesjährige Themenwahl ist natürlich ihr aktueller Bezug : Die Verwendung elektronischer Datenverarbeitung für die wissenschaftliche Dokumentation. Wie gross das Interesse an diesem Thema war, zeigte die hohe fachliche Eignung der Teilnehmer. Neben Spezialisten von Euratom und Teilnehmern aus etwa 10 Ländern des Gemeinsamen Marktes und der EFTA, nahmen Beobachter internationaler Organisationen und Einrichtungen, wie IAEO, ESRO, CETS, ENEA, UNESCO und CERN aktiv an den Diskussionen teil.

Die einzelnen Vorträge lieferten vollständige und ausführliche Auskünfte über alle Vorhaben der europäischen Gemeinschaft, deren Verwirklichung sich auf die Entwicklung der elektronischen wissenschaftlich-technischen Informationssysteme stützt. Die Vorträge gruppierten sich vor allem um vier Hauptthemen : Einführung mechanisierter Arbeitsmethoden in wissenschaftlichen Bibliotheken, automatische Schlagwortgebung als eine Ergänzung des halbautomatischen Dokumentationssystems, Verwendung, bei der wissenschaftlichen Dokumentation, elektronischer Hilfsmittel für Uebersetzungen — Systeme wie sie von CETIS, der Zentralstelle für die Verarbeitung wissenschaftlicher Informationen, entwickelt worden sind — und die Erfahrungen, die die ZID mit dem europäischen nuklearen Dokumentationssystem gewonnen hat.

Dankbar vermerkten die Organisatoren das rege Interesse aller Konferenzteilnehmer, das besonsers im Laufe der lebhaft geführten Diskussionen zum Ausdruck kam. Es rechtfertigt die Zusammenstellung der verschiedenen Referate, damit diese verteilt und einem grösseren Kreis von Interessenten zugänglich gemacht werden können.

Dies ist Sinn und Zweck der vorliegenden Veröffentlichung.

R. BREE

PREFACE

La réunion annuelle des bibliothécaires et documentalistes scientifiques de la Communauté Européenne travaillant dans le domaine nucléaire, est devenue une tradition, je dirais presqu'une institution. C'est en vérité une excellente tradition puisqu'elle permet, chaque année, aux responsables des bibliothèques scientifiques européennes d'échanger des informations sur les derniers progrès intervenus dans les domaines que les intéressent et sur leurs projets immédiats ou plus ou moins lointains.

La raison du choix de notre sujet de discussion pour cette année est évidemment son caractère d'actualité : j'ai nommé l'utilisation du traitement électronique des données en documentation scientifique. L'intérêt majeur qu'a suscité ce sujet a été illustré par la qualité des participants. Outre des spécialistes de l'Euratom et de quelque 10 pays du Marché Commun et de la zone de Libre Echange, un certain nombre d'observateurs appartenant à des institutions, des agences et des projets internationaux tels que l'AIEA, l'ESRO, l'ESTEC, l'ENEA, l'UNESCO et le CERN, ont participé activement aux discussions.

Les différents exposés qui ont été faits ont fourni des renseignements complets et détaillés sur tous les projets de la Communauté européenne dont la réalisation s'appuie sur le développement de systèmes d'information scientifique et technique mécanisés. Ces exposés touchaient quatre aspects principaux : l'introduction de la mécanisation dans les bibliothèques scientifiques, l'indexage automatique comme complément aux systèmes de documentation semi-automatique, l'application de la traduction automatique à la documentation scientifique tels qu'ils ont été dévelppés par le Centre de traitement de l'information scientifique (CETIS) — et l'expérience acquise par le CID — par le biais du fonctionnement du système européen de documentation nucléaire. L'intérêt de tous les participants, reflété par l'animation extraordinaire des périodes consacrées à la discussion, fort appréciée par les organisateurs, a justifié le rassemblement de ces différents rapports afin de les diffuser et de les offrir à un public plus large.

Ceci est le but de la présente publication.

R. BREE

PREFAZIONE

E' praticamente divenuto una tradiziona, se non proprio un'istituzione, che bibliotecari e documentalisti scientifici nucleari d'Europa si riuniscano una volta l'anno. Tale tradizione è lodevole, perchè in tal modo si offre loro ogni anno la possibilità di dare e ricevere informazioni sui più recenti sviluppi nei campi interessandoli particolarmente e sull'avviamento che verra seguito in un futuro più o meno prossimo.

La scelta dell'argomento di quest'anno trova la sua ragione nell'attualità dell'argomento stesso : uso del trattamento con calcolatori elettronici dell'informazione nella documentazione scientifica. Che questo argomento rivesta un carattere di elevato interesse generale è stato dimostrato dalla rosa dei partecipanti : oltre agli esperti dell'Euratom e a quelli rappresentanti una decina di paesi del Mercato Comune e dell'EFTA, hanno preso attivamente parte alle discussioni gli osservatori di istituzioni ed organismi internazionali come la IAEA, ESRO, ESTEC, ENEA, UNESCO e CERN.

Le diverse communicazioni presentate hanno fornito un quadro preciso e particolareggiato su tutti i progetti europei interessati allo sviluppo di sistemi di informazione scientifica e tecnica basati sull'uso di calcolatori elettronici. Quattro si sono rivelati essere i principali argomenti : adozione della meccanizzazione nelle biblioteche scientifiche ; la codificazione automatizzata a complemento dei sistemi semi-automatici di documentazione ; l'uso della traduzione automatica nella documentazione scientifica, quali sono stati sviluppati dal Centro di Trattazione dell'Informazione Scientifica — CETIS — ; l'esperienza acquisita dal CID col suo sistema europeo di documentazione nucleare — ENDS —. L'interesse dei partecipanti si è manifestato con una viva partecipazione alle discussioni e l'attualità dei loro interventi ha suggerito di farne una raccolta che potesse essere inviata non solo ai partecipanti ma anche a quanti vi fossero interessati.

Tale è, dunque, lo scopo della presente pubblicazione.

R. BREE

VOORWOORD

De jaarlijkse vergadering van Europese bibliothecarissen en documentalisten werkzaam op nucleair gebied is een traditie, ja bijna een instelling geworden. Hiermede wordt elk jaar aan de leiders van nucleaire bibliotheken de gelegenheid geboden in internationaal verband inlichtingen te verkrijgen en te verstrekken over datgene, waartoe de nieuwste ontwikkelingen op de voor hen van bijzonder belang zijnde gebieden geleid hebben, en wat de verwachtingen zijn voor de nabije en meer verre toekomst.

De reden voor de keuze van het onderwerp van dit jaar is natuurlijk zijn actualiteit, namelijk : het gebruik van electronische dataverwerking in de documentatie. Het belang van dit onderwerp wordt geïllustreerd door het hoog niveau van de deelnemers. Naast specialisten van Euratom en van een tiental landen behorende tot de Europese Gemeenschap en de EFTA, namen ook een aantal waarnemers van internationale instellingen, zoals IAEA, ESRO, ESTEC, ENEA, UNESCO, CERN, actief deel aan de discussies.

De verschillende voordrachten gaven een grondig en gedetailleerd inzicht in alle projecten van de Europese Gemeenschap, die zich bezig houden met de ontwikkeling van wetenschappelijke en technische informatie-systemen waarbij elektronische rekenmachines gebruikt worden. Zij waren gericht op vier hoofdonderwerpen : introductie van mechanisatie in wetenschappelijke bibliotheken, automatische indexering als een aanvulling op het door Euratom ontwikkelde semi-automatische documentatie systeem, het gebruik van automatische vertalingen in wetenschappelijke documentatie zoals ontwikkeld door het Wetenschappelijk Informatie Verwerkings-Centrum — CETIS — en de ondervinding door het CID opgedaan bij de exploitatie van het Europese Nucleaire Documentatie Systeem — ENDS.

De interesse van alle deelnemers aan de vergadering, die tot uiting kwam in de buitengewoon levendige discussies — waarvoor de organisatoren dankbaar zijn — maakte het noodzakelijk de verschillende voordrachten te verzamelen en ze beschikbaar te maken voor ruimere kring.

Dat is het doel van deze publikatie.

R. BREE

LIST OF PARTICIPANTS

R. ABBEL	Institut für Dokumentationswesen 6 Frankfurt/Main Vogtstraße 50 Germany
A. ALUFFI	C.N.R. Piazzale Scienze, 7 R o m a - 00100 Italy
M. BALLARIN	C.S.N. Casaccia Casella Postale 2400 R o m a - 00100 A.D. Italy
G.N. BARTEN	Reactor Centrum Nederland Westerduinweg 3 P e t t e n Holland
H.H. BERNSTEIN	Euratom C.C.R. C.E.T.I.S. - Library I s p r a - 21020 Italy
T. BERTOCCHINI	S.O.R.I.N. Saluggia - 13040 Italy
R. BREE	EURATOM - C.I.D. Luxemburg Gd-Duchy
G. BRICARELLI	Centro Sperimentale Metallurgico Via del Castel Romano R o m a Italy
G.D. BRIX	EURATOM - C.I.D. Luxemburg Gd-Duchy
E. BRONS	C.E.N. M o l Belgium
BROSSMANN	Gesellschaft für Kernforschung mbH. 75 Karlsruhe Postfach 3640 Germany
P.U. BRUSA	EURATOM C.C.R. C.E.T.I.S. - Library I s p r a - 21020 Italy
H. BUNTROCK	EURATOM C.C.R. C.E.T.I.S. - Library I s p r a - 21020 Italy

S. CAPOBIANCHI	EURATOM C.C.R. C.E.T.I.S. I s p r a - 21020 Italy
L. CAVARA	EURATOM - C.I.D. Luxemburg Gd-Duchy
H. COBLANS	I.A.E.A. W i e n 1010 Kaerntnerring 11 Austria
M. CURVALLE	C.E.A. S a c l a y - 91 France
G. DEL BIGIO	I.A.E.A. W i e n 1010 Kaerntnerring 11 Austria
DEROBERT	E.S.R.O. 114, avenue de Neuilly Neuilly-sur-Seine - 92 France
A. DOYLE	European Space Technology Centre Domeinweg Noordwijk Holland
J. EVRARD	C.E.A. B.P. No. 6 Fontenay-aux-Roses - 92 France
H. FANGMEYER	EURATOM C.C.R. C.E.T.I.S. I s p r a - 21020 Italy
E. FEHER	ETH-Bibliothek Z ü r i c h 8001 Leonhardstraße 33 Switzerland
E. FIAMOZZI	EURATOM C.C.R. C.E.T.I.S. I s p r a - 21020 Italy
J. FISCHER	Hahn-Meitner-Institut für Kernforschung Berlin 1 Berlin 39 Glienicker Straße 100 Germany
F. GRAMENZ	Kernforschungsanlage GmbH. J ü l i c h 517 Postfach 365 Germany
C. GROENEVELD	Technological University Eindhoven Holland
A. GUENTHER	C.E.R.N. G e n è v e 23 Switzerland

J.H. HASPERS	Bibliotheek Technische Hogeschool Doelenstraat 101 D e l f t	Holland
L. HESPEELS	C.E.N. M o l	Belgium
J.C. IN'T'VELD	Reactor Centrum Nederland Westerduinweg 3 P e t t e n	Holland
ISSAEV	I.A.E.A. W i e n 1010 Kaerntnerring 11	Austria
M. JUNG	C.E.A. Saclay - 91	France
KEMMERICH	Gesellschaft für Kernforschung mbH. 75 Karlsruhe Postfach 3640	Germany
M. KNIBIEHLER	C.E.A. Cadarache - 13	France
L. LAFUENTI	C.S.N. Casaccia Casella Postale 2400 R o m a - 00100	Italy
M. LEVIDELEC	EURATOM Bruxelles	Belgium
L. LUCIANI	C.I.S.E. Casella Postale 3986 Milano - 20100	Italy
G. LUSTIG	EURATOM C.C.R. C.E.T.I.S. I s p r a - 21020	Italy
MADER	E.S.R.O. 114, avenue de Neuilly Neuilly-sur-Seine - 92	France
U. MAYER	EURATOM C.C.R. C.E.T.I.S. I s p r a - 21020	Italy
K. MELLENTIN	D.E.S.Y. 2 Hamburg 52 Nofkestieg 1	Germany
K.H. MEYER-UHLENRIED	EURATOM C.C.R. C.E.T.I.S. I s p r a - 21020	Italy

O.A. MIKHAILOV	U.N.E.S.C.O. Place de Fontenoy Paris - 7e	France
E. MOEBIUS	EURATOM C.C.R. C.E.T.I.S. - Library I s p r a - 21020	Italy
E. MURA	EURATOM C.C.R. C.E.T.I.S. - Library I s p r a - 21020	Italy
F. PECO	EURATOM - Diffusion des Connaissances Luxemburg	Gd-Duchy
M. PERRET	E.N.E.A. 38, bld Suchet Paris -16e	France
S. PERSCHKE	EURATOM C.C.R. C.E.T.I.S. I s p r a - 21020	Italy
A. PETRUCCI	EURATOM C.C.R. C.E.T.I.S. I s p r a - 21020	Italy
G. POZZI	EURATOM C.C.R. C.E.T.I.S. I s p r a - 21020	Italy
W. RITTBERGER	Z.A.E.D. 6 Frankfurt/Main Westendstraße 21	Germany
L. ROLLING	EURATOM - C.I.D. Luxemburg	Gd-Duchy
J. ROSEN	E.N.E.A. EURATOM C.C.R. I s p r a - 21020	Italy
M. SAUBESTY	C.E.A. Cadarache - 13	France
M. SCHAUER	C.E.A. Grenoble - 38	France
K. SCHNEIDER	Zentralstelle für Maschinelle Dokumentation 6 Frankfurt/Main Holzhausenstraße 44	Germany
M. TISLJAR	Oesterreichische Studiengesellschaft für Atomenergie W i e n - 1082 Lenaugasse 10	Austria

L. TIKHONOV	I.A.E.A. W i e n - 1010 Kaerntnerring 11	Austria
V. TYURNIN	I.A.E.A. W i e n - 1010 Kaerntnerring 11	Austria
J.P. VAN TWIST	Twente Technological University P.O. Box 217 Enschede	Holland
W.M. VADEN	I.A.E.A. W i e n - 1010 Kaerntnerring 11	Austria
C.L. VAN DEN MUYZENBERG	EURATOM C.C.R. C.E.T.I.S. I s p r a - 21020	Italy
K. VEIEN	Danish Atomic Energy Commission Riso-Roskilde	Denmark
M. VERHOEF	EURATOM - C.I.D. Luxemburg	Gd-Duchy
C. VERNIMB	EURATOM - C.I.D. Luxemburg	Gd-Duchy
G. WENSKE	Z.A.E.D. 6 Frankfurt/Main Westendstraße 21	Germany
J.E. WOOLSTON	I.A.E.A. W i e n - 1010 Kaerntnerring 11	Austria

THE INTEGRATED SYSTEM OF THE ISPRA LIBRARY

by

W. Rittberger [x)]

ABSTRACT

A short description of the mechanized and integrated system is given.

Bibliographical and bibliotechnical details for the preparation of the data input and also the lists and catalogues for the integrated procedure are defined and illustrated.

x) Zentralstelle für Atomkernenergie-Dokumentation (ZAED)
Frankfurt am Main

Our library's integrated and mechanized library system was developed in collaboration between CETIS and the group responsible for technological and scientific information in Ispra, the so-called Library & Documentation Department.

This development was carried out as part of the research and development tasks assigned to CETIS and (I should like to stress this point) it was constantly encouraged and supported by the CETIS management, the Centre's directorate, and by the Directorate "Dissemination of Information" in Brussels.

The first attempts to prepare acquisition lists and bound catalogues by computer date back as far as 1962-1963; due to internal difficulties, however, they were not carried through at that time. Also, since that time, ordering, accounting and inventorying of journals have been handled by our computer facility. In addition, all filing and cataloguing work connected with publications of members of the Centre's staff has been handled, since that time, by means of an IBM 7090 computer.

In the second half of 1965 we considered extending the scope of mechanized ordering, accounting, and inventorying to cover other types of publications as well, such as monographs, series, etc. This automatically brought back into focus the problems connected with the computerized preparation of acquisition lists and bound catalogues. Unfortunately, or maybe fortunately, the existing programs prepared for the Centre's own publications and for journals would not do for the extended scheme. Following a thorough discussion, it was agreed to develop a computer-based integrated system to cover all of the library's operational process. The change from

7090 to 360 hardware anticipated at CETIS was a decisive factor in making that decision.

Why and when do you need a computer-based system? Much has been written and said on that subject with recurrent emphasis on the fact that the time-consuming routine operations in various sections of a library can be carried out by the computer. The answer to our question may be split into two parts; one partial answer covers the internal operation of a library, the other one takes care of the external relationship a library maintains with its users and of the dissemination of information on all documents of each other, of course.

As far as mechanization and rationalization of the internal operation is concerned, different points will be emphasized for different types of libraries. University libraries e.g. will primarily be concerned with improving their loan procedure or with mechanized preparation and updating of catalogues. A specialized library or a documentation department like ours should, in my opinion, assign priority to stepping up the processing routine for an ever increasing volume of input material and to making that material available to the user in a short time. This should of course imply that technical details, e.g. loan procedures, are handled in a way that is most efficient and most convenient for the user. As an example I mention the Euratom-CID information system which will be discussed tomorrow. In case Selective Dissemination of Information is offered to scientists, check-in of periodicals, requests for missing copies, and comprehensive lists of periodicals on hand (lists which the scientist should by all means have on his desk) will be of vital importance to a library.

As far as the external part, i.e. the relationship with the user, is concerned, the various kinds of machine-made acquisition lists, bound catalogues, the fast access to specialized catalogues covering specific fields, and SDI service drawing from all incoming publications provide for an essential improvement and expansion of the information services offered. You are all quite familiar with that point and you might agree that this development, which surpasses anything that could be expected from manual efforts, will for the first time establish a continuing and habitual relationship between the scientist and his information center.

As I mentioned above, work on systems analysis was begun in 1966. Every detail of a library's operational procedure, from acquisition through cataloguing, loan and information services, was discussed in collaboration with Mr. Capobianchi and Mr. Petrucci, who will later talk about all technical and programming details. Sometimes our discussions were quite lengthy because, nonwithstanding the severe restrictions imposed on the librarian's work by existing regulations, there is very much leeway for personal decisions in individual cases. When the analysis was completed it seemed to be feasible to eliminate all but a few manual filing procedures and other operations and to organize our library according to the following scheme.

The publications to be processed within the system are categorized as follows, according to their type and to the manner in which they are processed in the operational scheme.

M	monographs
MS	monographical series (publisher series)
P	periodicals
PS	periodical series
R	reports
OP	opus
T	thesis
PT	patents
PM	pamphlets
A	articles; analytics

We identified the following acquisition modes:

SING	single order
SUBS	subscription
STOR	standing order
SECT	section
COMP	compact order
MEMB	membership
FREE	free of charge
EXCH	exchange
LOAN	loan
BACK	backlog

It is evident that, as a consequence of the above-mentioned requirements necessary for the internal operation as well as for information services to the user, the output can meet all demands only if one provides for a very detailed input of data. Thus the following categories were identified for data input:

1	identification
2	classification
3	author(s)
4	title(s)
5	imprint
6	other bibliographical data
7,8	not yet occupied
9	administrative data

As a rule all data should be entered only once. In order to avoid repetitious work when making corrections or amendments, the statements listed within individual categories were themselves divided into fields. The input procedure will be demonstrated by some examples (fig. 1).

There are three facilities provided for data input:

1. punched paper tape;
2. punched cards;
3. IBM 2260 Display Unit.

Paper tape is suited for primary input, punched cards are suited for check-in e.g. of periodicals, and the IBM 2260 console adapts to on-line operation in check-in of accessions and in loan procedures.

The output provides for working with the following lists and catalogues:

A Bibliotechnical lists

Lists concerning acquisition work:

1. 1M = "On-Order"-list for monographs, monographical series, thesis, patents, etc.

   1P = "On-Order"-list for periodicals

   1PS = "On-Order"-list for periodical series

   1R = "On-Order"-list for reports

   lists 1M, 1P, 1PS arranged alphabetically, list 1R arranged by report-number.

2. 2A "On-Order"-list arranged by order-number
   2B "On-Order"-list arranged by dealer

   (these lists are only asked if the order cannot be identified by lists mentioned under 1.)

3. 3A List for orders to the dealer
   3B List for claims to the dealer

4. Lists announcing receipt of publication

   4P1 = list for periodicals arranged by title/service

   4P2 = list for periodicals arranged by service/title

   4M = list for all other publications on reservation, arranged by location-number.

5. Title-list of periodicals, kept within a department. This list is needed because part of the periodicals is on permanent loan at the different departments and each year the renewal of subscription has to be requested.

Check lists for accounting

1. List for settlement of invoices, arranged by dealer/order-number.
2. Cumulative list of the inventory.
3. List of publications not yet paid but already in possession of the library.

Lists for the loan-procedure

1. List of all publications on loan, arranged by location-number.
2. List of publications on permanent loan, arranged by department and/or user and alphabetically (catalogue of publications held within a department).
3. Claim-list to the users.

Library statistics

Various lists for library statistics are performed by the system.

B Catalogues

1. Main catalogue

   This catalogue is the main source for consultation by the clients and contains only the bibliographical details useful for this purpose, such as author's name, title of publication, location number and imprint.

It is arranged alphabetically according to author/title and title/author, i.e. the publication is entered in the catalogue in alphabetical order both under the author's name and under the title.

2. Subject catalogue

This catalogue is also a source for consultation and has the same bibliographical contents as the main catalogue.

It is arranged by subject and year of publication (the latter in decreasing order); if the publication covers several subjects it is entered under every one of the subjects concerned.

3. Main catalogue/shelf list

This catalogue contains the complete bibliographical and bibliotechnical data on each publication held by the library and recorded by the system. It is arranged according to the location number and is essential for the use by the library staff.

4. Periodicals catalogue

This catalogue contains all the bibliographical data including the holdings relating to the periodicals recorded by the system. It is arranged alphabetically according to the title of the periodical.

5. Conference catalogue

This catalogue is produced in three parts
a) by subject (KWIC), b) by place, c) by date.

6. Special catalogues

   For: monographical series; periodical series; corporate authors, patents and pamphlets.

7. Accession list

   This list contains all new publications arriving at the library within a fixed period.

8. Special information list for determined subjects

   Based on "customer-profiles", information on new arriving publications is distributed.

Some examples are given for bibliotechnical lists (fig.2). The examples for the individual catalogues will be explained to you in detail in this afternoon's discussion and demonstration session and there will be sample copies of these catalogues distributed to you.

In conclusion I should like to point out that the operational routines for loan and the statistical surveys to be made are also defined in detail. As yet no programming has been done in connection with loan procedures because the entire backlog must be cleared before the mechanized loan scheme can be put into operation. Besides, there is a wait until early 1969 before delivery of the IBM 2260. As far as statistics is concerned, we plan to have all the customary procedures of library statistics done by computer. In addition, however, we hope to extract from our loan statistics some information to be used in making our acquisition policy. At the same time, substantiated criteria for weeding out items and putting them to archival storage can

be established, e.g. if the space available to a library can no longer be expanded.

Acknowledgement is given to the help of Miss Möbius, Miss Zonca, and Mr.Brusa, during the developing of the system.

Figure 1 - examples for input

```
1       M10004
3N      Jacchia, Enrico
4I      (Il) Rischio da radiazioni nell'era nucleare. La tutela della
        popolazione e dei lavoratori, norme internazionali di protezione,...
5I      ; Milano - Giuffre
91      G; 5
932     108,3; 110,4; 400,5
934     REF,1; STO,2

1A      M10004; C5; MP570
2       Pg; Cb
5I      1963
5I1     607 p.
9I      G; 597/64; 040664; 5; LIT; 23000,00

1       M10004; Ld
3N      Jacchia, Enrico
4D      Atom, Sicherheit und Rechtsorderung
5I      1965; Freudenstadt - Lutzeyer
91      MB; 1
932     100,1

1A      M10004; Ld; C1; MP570
2       Pg; Cd
5I1     564 p.
9I      MB; 3802; 220965; 1; DM; 63,00

1       M10004; Lf
3N      Jacchia, Enrico
4F      Atome et securite. Le risque des radiations a
        l'age nucleaire
5I      ; Paris - Dalloz
91      V; 1
934     REF,1

1A      10004; Lf; C1; MP570
2       Pg; Cb
5I      1964
5I1     614 p.
9I      V; 007981; 150465; 1; FF; 54,00
```

fig.1a

```
1      MS7002
3N     Parsegian, V. L. (Ed)
4E     Nuclear science and technology
4ES    (A) Series of monographs and textbooks
5I     ; New York - Academic Pr.
91     A,BACK; 1
934    REF,1

1      MS7002; V1; C1; MM185
2      M1
3N     Flagg, John Ferard (Ed)
4E     Chemical processing of reactor fuels
5I     1961
5I1    XI,530 p.

1      MS7002; V2; C1; MM267
2      Mc
3N     Yeater, Max Laverne (Ed)
3C1    Rensselear Polytechnic Institute
3C2    American Nuclear Society
4E     Neutron physics
4EC    Proceedings of the symposium at Troy, N.Y., May 5-6, 1961
5I     1962
5I1    XIII,302 p.

1      MS7002; V3; C1; MM403
2      Mc
3N     Clark, Melville; Hansen, Kent F.
4E     Numerical methods of reactor analysis
5I     1964
5I1    XI,340 p.

1      MS7002; V4; C1; MM550
2      Mb; Nf3
3N     Haffner, James Wilson
4E     Radiation and shielding in space
5I     1967
5I1    IX,347 p.
```

fig.1b

```
1      P357
4E     IEEE Transactions on... (all sections)
5I     ; New York - IEEE
91     KUW,SECT; 1

1      P357.1; V16; N1
2      110
4E     IEEE Transactions on audio and electroacustics
5I     ; New York - IEEE
6H     1; 1; 14,,1966; ,,....
91     KUW,SECT; 1
92     0368; 0168; 40; 14
934    STO,1

1      P357.5; V15; N1
2      130
4E     IEEE Transactions on nuclear science
5I     ; New York - IEEE
6H     1; 1; 10,,1963; ,,....
6H     2; 1; 10,,1963; ,,....
6H     3; 1; 10,,1963; ,,....
6H     4; 1; 10,,1963; 12,,1965
6H     5; 1; 10,,1963; 12,,1965
6H     6; 1; 10,,1963; 11,,1964
91     KUW,SECT; 1
92     0268; 0168; 30; 1,6
934    CON,1
94     CON
```

fig.1c

```
1      P357.5; V15; N1
91     KUW; 2
934    311,2; 402,3

1      PS5007; V8
3N     Alt, Franz L. (Ed); Rubinoff, Morris (Ed)
4E     Advances in computers
5I     1967; New York - Academic Pr.
91     MB; 2
92     0867; 0167; 60; 1,1
934    REF,1; 510,2

1A     PS5007; V8; C2; MDY4
2      Df1; Bd
6H     1; 1; 1,,1960; ,,....
6H     2; 1; 1,,1960; ,,....
9I     MB; 7222005; 090867; 1; HFL; 57,70
9I     MB; 7222004; 090867; 1; HFL; 57,70
```

fig.1d

```
1      A5015; V44; MEY30
2      Eb
3N     Olsen, Haakon A.
4E     Applications of quantum electrodynamics
4EC    (Springer tracts in modern physics. Vol.44, pp.83-201)
5I     1968
```

fig.1e

```
1M    LIST OF MONOGRAPHS ON ORDER

1     M10004;Lf
3N    Jacchia, Enrico
4F  * Atome et securite. Le risque des radiations a l'age nucleaire
5I    ;Paris - Dalloz
91    V,SING;1
934   REF,1

1     M10010
3C1   Commissariat a l'Energie atomique
4F  * Colloque de metallurgie, 5ieme
4FC   Les gaz dans les metaux. Org. a Saclay les 26 et 27 juin 1967
5I    1962;Paris - Pr.univ.de France
91    V,SING;2
934   REF,1
934   520,2

1     M10012
3C1   Centre international d'Informations de Securite et d'Hygiene du Travail
4F  * Colloque international sur les accidents electriques, Paris, 2-5 mai 1962
5I    ;Geneve - CIS
91    CIS,SING;1

1     M18937
2     De;Df2
3N    Bowker, Albert H.
3N    Liebermann, Gerard J.
4E  * Engineering statistics
5I    ;Englewood Cliffs, N.J. - Prentice-Hall
5I1   XIV,585 p.
91    MB,SING;12
```

fig.2a

```
1P    LIST OF PERIODICALS ON SUBSCRIPTION

1     P317;V12;N1
2     040
4E  * IBM Journal of research and development
91    KUW,SUBS;1
92    0168;0168;30;1,6
934   MAT,1

1     P357.1;V16;N1
2     110
4E  * IEEE Transactions on audio and electroacustics
5I    ;New York - IEEE
91    KUW,SECT;1
92    0368;0168;40;14,1
934   STO,1

1     P357.5;V15;N1
2     130
4E  * IEEE Transactions on nuclear science
5I    ;New York - IEEE
91    KUW,SUBS;2
91    KUW,SECT;1
92    0268;0168;30;1,6
934   CON,1
934   311,2;402,3
94    CON

1     P357
4E  * IEEE Transactions on... (all sections)
5I    ;New York - IEEE
91    KUW,SECT;1
```

fig.2b

```
1PS   LIST OF PERIODICAL SERIES ON STANDING ORDER

1     PS5000;V6;N1
2     Ge3
3N    Drew, Thomas B. (Ed)
3N    Hoopes, John W. (Ed)
4E  * Advances in chemical engineering
5I    1966;New York - Academic Pr.
91    A,STOR;1
92    1266;0166;60;1,999
934   REF,1

1     PS5007;V8;N1
2     Df1;Bd
3N    Alt, Franz L. (Ed)
3N    Rubinoff, Morris (Ed)
4E  * Advances in computers
5I    1967;New York - Academic Pr.
91    MB,STOR;2
92    0867;0167;60;1,1
934   REF,1
934   510,2

1     PS5006;V7;N1
2     Pf
3N    Elliott, Henry Wood (Ed)
4F  * Annual review of pharmacology
5I    1967;Palo Alto, California - Annual Reviews
91    MB,STOR;1
92    0567;0167;60;1,1
934   509,1
```

fig.2c

EURATOM - CCR CENTRO EUROPEO DI TRATTAMENTO DELL INFORMAZIONE SCIENTIFICA - BIBLIOTECA - 21020 ISPRA - VARESE -ITALIA

3A ORDER LIST

DEALER:
GOERLICH E C.
VIA SAN SENATORE 6/2
20122 MILANO

PLEASE, ENTER CUR ORDER AS FOLLOWING:

ORDER NUMBER M 10004 COPIES 05 SING

Jacchia, Enrico

(Il) Rischio da radiazioni nell'era nucleare. La tutela della popolazione e dei lavoratori, norme internazionali di protezione,....

Milano - Giuffre

PLEASE, SEND INVOICE IN 3 COPIES AND ALWAYS REFER TO OUR ORDER NUMBER.

fig.2d

EURATOM - CCR CENTRO EUROPEO DI TRATTAMENTO DELL INFORMAZIONE SCIENTIFICA - BIBLIOTECA - 21020 ISPRA - VARESE -ITALIA

3A ORDER LIST

DEALER:
KUNST UND WISSEN - ERICH BIEBER
POSTFACH 46
D 7 STUTTGART 1 / GERMANIA

PLEASE, ENTER OUR ORDER AS FOLLOWING:

ORDER NUMBER P 357 COPIES 01 SECT
IEEE Transactions on... (all sections)
New York - IEEE

PLEASE, SEND INVOICE IN 3 COPIES AND ALWAYS REFER TO OUR ORDER NUMBER.

fig.2e

EURATOM - CCR CENTRO EUROPEO DI TRATTAMENTO DELL INFORMAZIONE SCIENTIFICA - BIBLIOTECA 21020 ISPRA - VARESE -ITALIA

3B CLAIM LIST

DEALER:
MAYER'SCHE BUCHANDLUNG - M. FALTER
BUCHKREMERSTR. 5
D 51 AACHEN / GERMANIA

WE ARE CLAIMING:

ORDER NUMBER Z 30
Atompraxis
COPIES 01 VOL. 00014 Y 68 NR 00002

ORDER NUMBER Z 159
Recueil des travaux chimiques des Pays-Bas
COPIES 01 VOL. 00087 Y 68 NR 00001

ORDER NUMBER Z 159
Recueil des travaux chimiques des Pays-Bas
COPIES 01 VOL. 00087 Y 68 NR 00002

ORDER NUMBER Z 10
Werkstoffe und Korrosion
COPIES 01 VOL. 00019 Y 68 NR 00001

PLEASE, CONFIRM OUR CLAIM.

fig.2f

4P1 PERIODICALS RECEIVED ON

NEW TITLES	PERM. LOAN	
	311	Y68 V00037 N00001 - Z 678 Alta frequenza
	311	Y68 V00037 N00002 - Z 678 Alta frequenza
	STO	Y68 V00058 N00001 - Z 339 Annali di chimica
	STO	Y68 V00058 N00002 - Z 339 Annali di chimica
	MAT	Y68 V00087 N00001 - Z 1264 Annals of mathematics
CON	CON 650	Y68 V00014 N00001 - Z 30 Atompraxis
CON	CON	Y68 V00014 N00002 - Z 30 Atompraxis
	108	Y68 V00046 N00001 - Z 1190 Klinische Wochenschrift
	108	Y68 V00046 N00002 - Z 1190 Klinische Wochenschrift
	108	Y68 V00046 N00003 - Z 1190 Klinische Wochenschrift
	108	Y68 V00046 N00004 - Z 1190 Klinische Wochenschrift
	108	Y68 V00046 N00005 - Z 1190 Klinische Wochenschrift
	108	Y68 V00046 N00006 - Z 1190 Klinische Wochenschrift
	520	Y68 V00060 N00001 - Z 1358 Metallurgia italiana
	STO	Y68 V00087 N00003 - Z 159 Recueil des travaux chimiques des Pays-Bas
	STO	Y68 V00019 N00002 - Z 10

FIG! 2g

4P2 PERIODICALS FOR PERMANENT LOAN TO:

CON BIBLIOTHEQUE - CONSULTATION

Z 30 - L Y68 V00014 N00001 EX01 Atompraxis
CON BIBLIOTHEQUE - CONSULTATION

Z 30 - L Y68 V00014 N00002 EX01 Atompraxis

fig. 2h

Computers of the Third Generation and the Library of the Ispra Center

S. Capobianchi

The problem of automating library functions is a problem of general interest. Studies and experiences are made everywhere, and CETIS, particularly interested by its own vocation in Scientific Information Processing, can not stay aside.

The task of CETIS is in fact twofold : to solve by means of computers the problems posed by various "customers" and, at the same time, contribute to the investigation and the development of new data processing techniques.

The system we have studied and which we describe here in its fundamental ideas and general outlines, is adapted to fully comply with the requirements of the Ispra library. Moreover it is, in our opinion, a pilot experience useful also to other scientific libraries which may wholly or partly use this system or at least extract ideas and solutions and adapt hem to their own specific needs.

We tried to realize a system applicable to various cases, being as general as possible. We believe to have thus contributed, to some extent, by showing the wqy (among the many possible) how to uniform and standardize the great task of library automation, and we do hope that our efforts will be useful also to other librarians and system designers.

The concept of Integrated Systems

First of all it seems necessary to clarify the meaning of the definition "integrated system" given to the process we have programmed.

By "integrated system" is meant, according to the terminology in use, an automatic system in which :

- the computer takes into account the existing relations between the different procedures constituting the system
- the different correlated procedures are coexisting in the memory of the computer and used on the basis of logical decisions taken by the system
- an adequate network of information, common to the various parts of the system, is continously at the disposal of the system
- each data is entered only once into the computer, even if it interests different procedures, and is treated at the same time for the different sectors interested by the data itself
- at the end of the elaboration process to which each item is submitted, the system disposes of a totality of completely updated data which may be used in the following elaboration procedure.

As it can easily be seen, this method of "data processing" deviates in a somewhat revolutionary manner from the -let us say- "traditional" concept of processing data in the sequential mode.

Until some time ago, in fact, the classical method of processing data connected to a certain administrative activity consisted in dividing the various functions to be carried out into separate procedures, having them elaborated and singularly controlled by the computer, and (eventually) collecting the data or results in a sort of synthesis program in order to obtain, by phases necessarily sequential in time, the desired results.

It is unnecessary to spend much time on illustrating the advantages an integrated system offers in comparison with a traditional one ; be it sufficient to mention the two most important ones :

- introduction of the data into the system only once, with consequent unification of the controls to be executed during the input phase, gain of computing time, etc...

- possibility of having at any moment an updated situation, relative to the whole system, for the totality of data introduced, up to that moment, into the computer.

The use of interrogation and control devices installed near the system user, but connected to the main computer will permit to profit to a larger extent by this constant availability of the system thus allowing, at any moment, to evaluate the real situation relative to a certain instant and to dispose "in real time" of all the informations controlled by the system and establishing a mode of communication of the "conversational" type between man an machine.

Once understood the interest of each organization effecting data management to a considerable extent, in the use of integrated systems one may ask, of course, why only now the realization of these methods is discussed nearly everywhere.

Several replies may be given. The two main reasons which delayed until 1966/67 the "boom" of the integrated systems are, in our opinion, the following :

- 1) Before computers of the third generation have been available there was, with some exceptions, practically no possibility of applying and utilizing correspondingly the philosophy set up by integrated systems. Though this philosophy was already known in a theoretical sense, it was not easy to realize.

  By means of the computers of the third generation have been put at disposal, and will more do so in the future, a "hardware" and a "software" which can be used in the desired sense.

  Hardware

  With regard to "hardware" (that is the whole of the electronical and mechanical devices by which a computer is composed), I wish to draw the attention particularly on the auxiliary direct access memories.

  As may be known, a memory is called auxiliary when it is not the core storage which is an integrated part of the computer where mainly the programs are stored. The auxiliary memories mentioned above are

characterized by having direct access facilities. This distinguishes them from memories of the sequential type, where the information can be obtained only by sequential explorations of the stored data.

As its name indicates, the direct access memory allows to obtain by means of direct address any data without having to scan, beforehand, the material being of no use with respect to the searched data item.

Among the machines put on the market by IBM, there are today disk memories, type 2314, where about 29 millions of alphanumerical characters can be stored in each pile of disks ; the access time is in the order of some 10 milliseconds. Each type 2314 consists of 8 disk packages which can be used simultaneously totaling to about 233 millions of bytes. Each physical disk pile can of course be replaced by another one of the same characteristics.

"Leaf" memories called "Data Cell" are also offered on the market which can store up to 400 millions of alphanumeric characters with relatively low encumbrance and costs, and with an access time of about half a second. There exist also magnetic drums (type 2301) on which it is possible to record programs or data up to 4 millions of alphanumeric bytes accessible in less than 10 milliseconds and transferrable to the central memory at a speed of 1.200.000 characters/sec. Not to mention the auxiliary memory modules (LCS) connectable to the central memory (modules of 1 or 2 millions of characters) where programs or data tables may be stored and transferred to the central memory at a speed of the order of a microsecond.

However, the tendency towards auxiliary memories of greater capacity, of easier and quicker access, and of lower price is practically only at its beginning.

Still in the "hardware field", IBM offers terminal devices of different capacity and nature which can be connected to the computer, such as teletypes, card readers, punched tape readers, printed, video equipments etc.

It must be emphasized how by the availability of "buffers" assigned to this types of equipment it is able to run them by only occupying minimum fractions of the elaboration time of the computer. In the case of a video, for instance, the image can be maintained on the screen or can be written thereon - if we may say so - by typing on the special keyboard without occupying the central computer.

As a consequence : the possibility for the central computer, to dispose of hundreds of millions of characters always accessible and to obtain, to question, or to modify these characters by remote operation on terminals, is a reality today.

Software

Regarding the software part (by software we mean all the programs of general utility allowing the relatively simple and easy handling of devices which constitute the entire computing installation), also in this sector considerable progress has been made by computer manufacturers. Above all, the possibility of working in "multiprogramming" definetely opened the way to the utilization of integrated systems, processing data in real time. By

"real time processing" is understood the possibility of interfering practically instantaneously with the normal working process of the computer in order to have effectuated an operation cycle which is, for instance, determined by a remote terminal.

Multiprogramming offers the possibility for the computer to treat and execute more programs simultaneously.
A superprogram governs and controls the work of the computer. This superprogram, known as "operating system", directs and distributes the calculation capacity of the computer among the different programs coexisting in the central memory. The long waiting times caused by the reading or transferring of information from or to the external auxiliary memory, are used by the "operating system" to elaborate other programs at the same time.

Still within the range of the possibilities offered by the "software" of the IBM computers of the third generation must be mentioned the new storage methods of data on direct access memories (especially "indexed sequential") which are particularly suitable for bibliographical data.

This does not mean that all difficulties would disappear by using "real time" working methods ; evidently special programs must be written and parts of the software prepared or adapted to specific needs. Technically, however, the way has been opened and the choice of one working method or another has become more liberal.

- 2) Another fact, in our point of view of no less importance, which opposed and in some cases still opposes to choose integrated working methods for the solution of the problems, is a psychological one and has sometimes its origin in the organization itself to which the system designer belongs.

To adopt an integrated system practically means not only to substitute certain phases of clerical work by analogous phases of automatic work, eventually checking phase by phase the partial results obtained, but very often to revolutionize the basic concepts of manual work and to reconstruct its entire organization. It means to identify in collaboration with the system designer the essential elements of the field of work to be automated, eventually reorganize it with regard to the new possibilities offered by the automation, and to coordinate the new tasks the clerical staff will be confronted with.

As regards the Ispra library itself, all pre-conditions required for the realization of a "total system", have fortunately been fulfilled.

Application of the Ispra Library

CETIS has at its disposal an IBM 360/65 computer, equipped with 512 K bytes of central memory. Disk storage units of the largest type (i.e. model 2314) are to be installed before August 1968, whilst the arrival of a Data Cell type memory (2321 type) (or of another 2314 disk storage unit) and of 4 video units (display units 2260) is planned for February next year.

A promotor in mechanizing his service and a most valuable collaborator has been Mr. Rittberger, head of the Ispra library(*) ; this was of fundamental importance.

There was thus no obstacle against the realization at the library of an integrated system operating in "real time".
It was motivated by the following reasons :
The main functions of the library are
a) acquisition
b) cataloging
c) loans
d) information

As known, all problems of each of the above functions are interdependent ; certain data required for acquisition are as well useful for cataloging, some data necessary for loans are practically the same as for acquisition and cataloging, and so on.

It is, therefore, desired to introduce the data, whenever possible, only once into the computer and to have them then automatically distributed and treated according to their nature.
Moreover, in a library, operations must be carried out which require the consultation and updating of "files", the manual control of informations already existing, the reply as quick as possible to specific requests, the eventual correction of financial or administrational documents, etc.

The suppression of all manually used "files" is always possible by mechanization, but the difficulty is how to likewise provide the user with the information he needs. In the traditional way the "files" are substituted by mechanographical listings. This system serves in much cases, but not always. The main objection generally made to such a solution consists in the difficulty of continously updating these listings, especially in cases where the material is too voluminous to be completely re-printed repeatedly.

The problem is perfectly solved by the possibility to have the "files" required for administration pruposes stored on direct access magnetic memories. In this way the data is always available by terminal devices, thus allowing the storage, control and automatical updating, and the furnishing upon request of the same complete information which can be found on a card kept in a file, with the only difference that this is a perfect file in which nobody can put disorder and where the information may be arbitrarily multiplied in order to be utilized wherever required.

In the case of the Ispra library, we thought it useful, to begin with, to organize "in real time" (by means of display units) all the functions related to loans and arrivals. Other possibilities will be studied in the future when after the installation of these devices we shall have some practical experience.

I think to have sufficiently explained the reasons which determined us to choose an "integrated system" by which all bibliographic as well as administrative informations may be collected.

(*) Present address : Zentralstelle fuer Atomkernenergie-Dokumentation (ZAED), Frankfurt, M., Germany. Mr. H.H. Bernstein, head of the library since February 1968, is also extremely engaged in realizing the library mechanization program.

Even if our system as a method of file storage and as a structure of programs is mainly based on the possibility of elaborating "in real time" the main operations of the library, we foresee to run the program "off-line".

It was necessary, furthermore, to plan a suitable emergency procedure in case the "display units" are "out of service". Besides, it is by means of this very procedure that we are able to test the system till these units will be installed.

It is to be noted here that certain parts of the system are not programmed to work in real time. One can, for instance, not imagine that the printing process for the catalogs is carried out immediately after a simple direct request by means of terminals ; this would also be totally uneconomic.

The functions to be executed "in real time" must comply with precise demands :
- be short in time, occupy the computer memory for a minimum time (in estimation a maximum of a few seconds) in order to leave the possibility to other users to use the memory modules reserved for the terminals

- refer to informations needing immediate and/or visual controls before being introduced in the system (checking of incoming material)

- refer to informations which have to be continually updated and available in order to be used and/or demanded at each moment (lending procedure).

All informations and all administrative functions for which it is not necessary to immediately dispose of the "computer", are not treated by the integrated system in "real time" ; typical examples are the input of orders for new publications, the printing of bibliotechnical lists ans catalogs, operations which may be executed at pre-fixed intervals.

Considerations about memory size, processing times and economics

At this point the following objections may be made : the described system disposes of a capacity of a 360/65 computer, of video terminals and imposing auxiliary memories : how can it be adapted to computing centers of far less capacity and not equipped with these devices ?

In fact, the system can also be used by computers of less capacity. First of all we want to point out that (except obviously for the possibility of direct interrogation) our procedure may also be used off-line. This is what we are doing at present. The organization of the files and of the programs is essentially based upon the availability of auxiliary direct access memories (less expensive than magnetic tapes), and can also be used without remote terminals.

The most extensive program requires at present 120 K bytes, but may be reduced to 80 K bytes by an overlap operation, i.e. re-reading of parts of the program in the central memory.

Thus also with aspect to the quantity of the needed memory, these programs are nevertheless feasible. As an indication we may consider that a computer of average size (IBM 360/40 type) generally has a memory of 128 K bytes, whilst a computer of the 360/50 type normally disposes of a memory of about 256 K bytes.

As regards the direct access memories, their capacity varies according to the dimensions of the library to be mechanized and corresponding to the number of publications to be handled.

In our case we shall probably be able to memorize all the files (except those containig the catalogs) on one (or one and a half) of the 8 Disk Packs 2316 constituting the disk unit 2314.

The catalog files (for the moment we plan 11 different catalogs) will be contained in 120-150 millions of bytes. The problem of cataloging has been solved by keeping into account a certain periodical (rather low) extraction of updated catalogs, whilst during the intermediate periods supplementary catalogs will be extracted.

Considering the necessary machine times, the costs for memorization and processing, the operational practicability, etc., we decided also for the catalogs to effectuate the updating by means of direct access memories rather than by merging the magnetic tapes (according to traditional methods). By elementarily modifying the programs, it is, however, possible to use magnetic tapes for the cataloging and to execute the updating by the standard merging and total re-writing of the data.

Finally, the programs which will use the terminals (we plan to use 2 display units 2260))will be contained in 40 K bytes (in the present case they will be stored on a "storage drum").

In conclusion, a library which disposes even for short weekly periods of a computer of average capacity may easily utilize our system. Though obviously one might not have fully made use of some particular advantages of the system, it will as well constitute a valid and efficient instrument.

The system is now in its operational phase with regard to the periodical publications and in preparation with regard to the other types of publications.

We guess that those parts of the system which will function on-line will be in operation in May 1969, whilst with regard to a quick documentation (SDI : Selective Dissemination of Information) it is impossible to predict an exact date.

We estimate the running costs of the system at the Ispra Center to about 3.000 $/month.

The new system will, when completed, be available at the Program Library of CETIS.

We remain at disposal for any further informations and are ready to collaborate with all persons interested, in order to study eventual adaptations of the system to specific demands.

The Integrated Library System - Programming Considerations

A. Petrucci

Following on the outline of our system by Dr. Rittberger, who dealt mainly with its "bibliotechnical" aspects, and the introduction by Dr. Capobianchi, who explained the general media and working methods on which the organization of the system is based, it is now my task to describe in greater detail how the various problems of automating the management of a scientific library are handled in practice - "integrated" automation, in the sense de-defined earlier by Dr. Capobianchi.

Naturally I cannot recount all the individual problems we had to consider and solve ; for one thing, our time is rather limited, and for another, I might bore most of you with highly technical arguments that only really interest a programme analyst. For that reason I shall confine myself to those aspects which, in my opinion, are of general interest to everyone who has already dealt, or expects to deal in the future, with the problems connected with the mechanization of the management of a scientific library.

INPUT CONSIDERATIONS

As we have already seen, library management comprises four basic "activities" or "functions" :

Acquisition
Cataloguing
Loan
Information

Because of the various nature of the publications handled by a library, and therefore the variety of management problems occasioned by that various nature, we broke down the publications into the following groups or "types" of publication :

M Monographs
MS Monographical Series
P Periodicals
PS Periodical Series
R Reports
OP Opus
T Thesis
PT Patents
PM Pamphlets
(A Articles)

Each one of these types requires special processing, not needed for the other types, at the "acquisition" stage (for instance, the procedure for ordering, checking in and paying for a periodical is obviously not the same as for a monograph) and at the "cataloguing" stage (various rules for distribution of the bibliographical data in the various catalogues ; special catalogues required for P, PS, MS, etc.)

Most of the types of publication envisaged by us are conventional and need no comment ; the "Article" type, shown in brackets (see above), is not a real publication type ; it was put in as a cross-reference, to enable the titles of particularly interesting articles, published in periodicals and periodical

series or reports, to be included in the catalogue ; or to permit of entries of subtitles of other publications, which could not have been shown in the catalogue in the ordinary way. By this means part of the "information" function mentioned earlier is effected at the "cataloguing" stage. Naturally, the "Article" publication type always refers to the LOCATION NUMBER of the publication from which it derives.
The OPUS type comprises publications of monographical character divided into several volumes, such as encyclopedias, technical handbooks, etc.

The data concerning each publication are fed to the system via the following catagories :

1. Identification
2. Classification
3. Author
4. Title
5. Imprint
6. Other bibliographical data
7,8. (not yet occupied)
9. Administrational data

Within each "category" the data are divided into "sub-categories" and then into "groups", so that the different characteristics of any one piece of information supplied to the system can be specified and automatically identified by the program.

The quantity of data relating to each bibliographical unit is extremely variable, ranging from a minimum of a few hundred characters to a maximum of about 4,000 characters. Moreover, the information relating to each catagory is of variable length, consisting of one or more "fields" which thermselves vary in length. Hence, if all the data had to be compiled and punched on rigidly predetermined lines, the task of preparing them would be impossibly complex, with respect to cost, time, space, error probability, etc. We therefore preferred to unload as many of the difficulties as possible onto the programming, so as, to keep the data compilation simple and practical and limit the punching work to strict essentials.

Every field belonging to a given category is separated from the next field by a ";" which operates as a field separator ; a missing field is shown by the presence of the field separator alone. Thus there is no predetermined position for any of the fields within a category ; and the length of each individual field is limited to the significant characters present. The alignment and expansion of the fields in fixed formats are performed by the programme by means of tables which specify the characteristics of the fields of each category.

The only exceptions to these rules are found in category 1, where :

- the first field, consisting of the ORDER NUMBER, must always be present,

- the successive fields are additionally identified by a specification code.

So it is not necessary to use the field separators when some fields are missing.

The reason why the ORDER NUMBER must always be present is that it was

decided to use as identification code for any publication (same title and author, within one of the types defined above) an articulated code based precisely on the ORDER NUMBER used at the time of acquisition of the first copy of that publication. The various editions, volumes, parts, supplements and numbers that may exist are distinguished by adding on to the order number the number of the edition, volume, part, etc. The ORDER NUMBER, used as the main component of the identification code in the sense just mentioned, is sufficient by itself to identify unequivocally the bibliographical units relating to the Periodicals and the Periodical Series (the details of VOLUME/ PART/NUMBER, etc., are recorded as HOLDINGS on the same unit). For the Monograph Series it suffices to add on to the ORDER NUMBER the number of the volume and/or part to distinguish different monographs of the same series. For the Monographs the identification code need only be extended beyond the ORDER NUMBER if the publication is physically divided into several volumes and/or parts.

Another point is that by using the ORDER NUMBER as the basic identification code we avoid having to form a new code simply for that specific function, and also avoid the need to establish a correspondence between the various ORDER NUMBERS used to order consecutive copies of any one publication on one hand, and the identification code for the copies on the other hand.

To obviate the introduction of other codes, the identification code must also be given the signification of INVENTORY NUMBER.

The figures 1 shows the list of the varioys categories and their respective functions.

The figures 2 to 7 show the contents of each field in each category. The maximum length of each field is given in parenthesis under each field. The field-separators ";" are also indicated.

To close the subject as regards the input of data into the system, I will just remark that the punching is done by means of paper tape typewriters. The paper tape is converted directly into magnetic tape by means of a paper reader connected to the computer. The use of paper tape typewriters enabled us to benefit from certain special charecters and upper-case and lower-case letters in the punching process ; the characters are printed on the output lists by a special printing chain.

As regards acquisition, we planned for the following "modes" :

SING	Single order	(M)
SUBS	Subscription	(P)
STOR	Standing order	(PS, MS, OP)
SECT	Section	(P)
COMP	Compact order	(all)
MEMB	Menbership	(all)
FREE	Free of charge	(all)
EXCH	Exchange	(all)
LOAN	Loan	(all)
BACK	Backlog	(all)

For each type of publication there is a standard mode of acquisition ; certain modes can be utilized for all the types ; it is also possible that a certain type of publication may be acquirec by a non-standard mode, e.g., back numbers of a periodical, or individual volumes of a set of monographs or of periodicals,

are acquired by a SINGLE ORDER ; monographs with a series of updating notes or news-letters are acquired by a STANDING ORDER or SUBSCRIPTION. Each combination of "Type of publication/mode of acquisition" has a particular checking procedure both for arrivals and for payments.

For "periodical/SUBSCRIPTION", each separate number is checked in (in the manual procedure by cardex), whilst a single payment covers the set. For "monographical series/STANDING ORDER" and for "periodical series/STANDING ORDER", both the arrival and the payment of each separate volume is checked.

In the case of "monograph + up-dating addenda/SUBSCRIPTION", a single payment covers the set, but the monograph and each of the addenda must be checked in separately ; the ORDER cannot be regarded as "filled" when only the monograph has arrived, but must be considered as partial delivery until the scheduled addenda have arrived.

I am not going further into these details, which I have only mentioned to illustrate some of the reasons for dividing the publications into different "types".

OUTPUT

The different catalogues and lists have been presented by Dr. Rittberger.

PROCESSING

Let us now turn to a brief explanation of the general organisation of the system.

Files

In the first place, I would draw your attention to the fact that when this system comes into operation, all types of Library card files used in the manual methods can be put aside.

Magnetic files, containing the necessary data, will then be at the system's disposal on the direct-access auxiliary storages. These files are all organized on the indexed sequential storage basis and are thus accessible in real time. The principal files are listed in Fig. 8.

Further details on the structure of the MASTER and ORDER files, which are the most interesting from the point of view of programming, can be provided on request.

Processing of bibliographical and bibliotechnical data

Regarding the general organization of the programmes, the figure 9 shows the system flow-chart.

Pre-cataloguing. The bibliographical details concerning each publication are introduced into the system at the moment of ordering the first copy of each one ; these data are used by the system to make up the ORDER to the dealer and are stored to form a base for the future cataloguing.

Bibliographical check and classification. When the publication arrives the pre-recorded data are modified (if necessary) and completed by the addition of the LOCATION NUMBER, the "subject" codes and any other bibliographical information not available at the time of ordering.

Backlog . The bibliographical details of the publications already owned by the library at the time when the system comes into operation are entered by means of "fictitious" orders ; in this way, from the procedural point of view at any rate, there is no substantial difference between the input of the publications already possessed by the library and the new publications acquired.

Filing and retrieval. The bibliographical data are recorded in the order of "identification codes" on a single "MASTER FILE", from which are obtained the bibliographical data needed to perform the various operations the system is designed for. For example, in the case of ORDERS for supplementary copies for a publication, the bibliographical data needed to compile the ORDER are automatically extracted from this file.

The system also has a DEALERS' FILE, from which are extracted the name and address of the dealer to whom the ORDER is to be sent. When the ORDER concerns a publication intended for permanent loan, there is also an arrangement for storing the "reservation" and signalling this fact when the publication arrives.

New publications "on order" but not yet arrived, which are recorded in the MASTER FILE, have a "pre-cataloguing" code and are not entered in any of the scheduled catalogues.

All orders are recorded in an ORDER FILE, used to check arrivals and invoices and, when necessary, to emit claims for orders not executed within the due time-limit. This file is particularly useful for checking the arrival of volumes of the Periodical and Monograph Series (standing orders) and especially for the periodical issues, where it is of prime importance that the claims be filled punctually. The data acquired and used for checking arrivals provide an exact up-to-date picture of the holdings in periodicals.

The data concerning the arrival of publications are transferred to the computer via the KEYBOARD of the IBM 2260 DISPLAY UNIT, which, installed at the library's "Reception" office, is connected ON-LINE to the computer. The data already in the system which relate to the newly-arrived publication are shown on the video screen, permitting a direct instant check on the correctness of the up-dating operation in progress. If the check is favorable, the new data are introduced into the system. In this way a CONVERSATIONAL exchange of information is established between the library staff and the system.

These "conversations" are recorded on the MISCELLANEOUS FILE. This file also records all messages, error messages, and bibliotechnical lists from the OFF LINE part of the procedure ; each record is characterized by a "record type" code and the relevant "sort key" is shown. The file is then arranged according to the "record type" and "sort key" and supplies the different lists and messages required.

Extraction of catalogues

The data for the various catalogues are extracted once and for all from the MASTER FILE in a single phase ; at the moment of this extraction a "catalogue code" is added and the appropriate "sort key" is shown. The bibliographical unit is repeated as often as necessary for the different catalogues. The FILE thus obtained is arranged according to the "catalogue code" and the "sort key" and forms the "CATALOGUES FILE", from which the various catalogues are printed. When the MASTER FILE is modified by the addition of new publications and/or by variation of the bibliographical data from the data already contained in the CATALOGUES FILE, the modified bibliographical units alone are extracted to form a "supplement" to the CATALOGUES FILE, on the SUPPLEMENTS FILE, while at the same time the CATALOGUES FILE is brought up to date. This supplement can be printed separately (thus supplying the supplements to the various catalogues).

Because of the huge quantity of data to be dealt with in the catalogues (about 150 million characters for about 20,000 different titles), this method of operating, although it demands a large auxiliary storage capacity, is certainly more convenient than extracting the data for the different catalogues from the MASTER FILE every time and arranging them every time, according to the various types of catalogue.

Circulation. The data necessary for the loan procedure are recorded on the LOAN FILE, which is used to check both "temporary" and "permanent" loans and also contains "reservations" ; the latter are automatically displayed when the "reserved" publication is available for loan.

Provision has also been made for the possibility that a publication on "permanent" loan may be temporarily taken back by the library and loaned out "temporarily" to another user ; on return from the temporary loan it is automarically assigned on permanent loan to the original user.

If a publication is not returned within the due time limit, a claim for it is emitted automatically.

The data needed for the loan are fed into the system via the KEYBOARD of the IBM 2260 VIDEO UNIT installed at the library's "Circulation" office. All the data concerning the requester and the publication requested are immediately displayed on the video screen for the purpose of checking them before the operation is performed. In this case too, as in the check-in, a "conversational" type of information exchange occurs between the library staff and the system.

To establish all the lists and messages concerning "circulation", the data contained on the LOAN FILE will be completed by the addition of the TITLE/ AUTHOR of the publication and the name of the user, extracted from the MASTER FILE and the USERS FILE respectively.

Invoices. The prepared invoices are recorded on the INVOICE FILE, which is the basis for up-dating of the INVENTORY FILE, the list of invoices for payment, and, on request, the up-dated inventory list.

Of course, this contribution is not, and cannot be, a complete one.
I would be glad, however, to give more detailed informations to everybody who wants them.

CATEGORIES

1	Identification for ordering or modifications
1A	" " receipts
1AU	" " " of updating addenda
1B	" " bibliographical information notice
1C	" " claim notes to dealer
11L	Linkage for unnumbered series
11R	Cross references
12	Location number of depending library
13	External-reference-number: Reports, Patents
2	Subject classification
22	" " of depending library
2KL	Key-words
3N	Nominative author(s)
3C	Corporate "
4L	Title
4LS	Subtitle
4LC	"
4LN	New title (Title change for P and PS)
5I	Imprint
5I1	"
6B	Bibliographical data
6C	Conference data
6H	Holdings of P and PS
91	Order
911	Order number to be referred to for compact order
92	Frequency
931/4	Reservation
935	Circulation
94	Notification for "Table of contents"
951/4	Loan
961/4	Return
9A	Receipt of publication (supplementary informations)
9I	Data of invoice
AD1/3	Dealer description
ADP	Department/service description
AU	User description
AS/P	Subject description
ARN/E	General references

Fig. 1

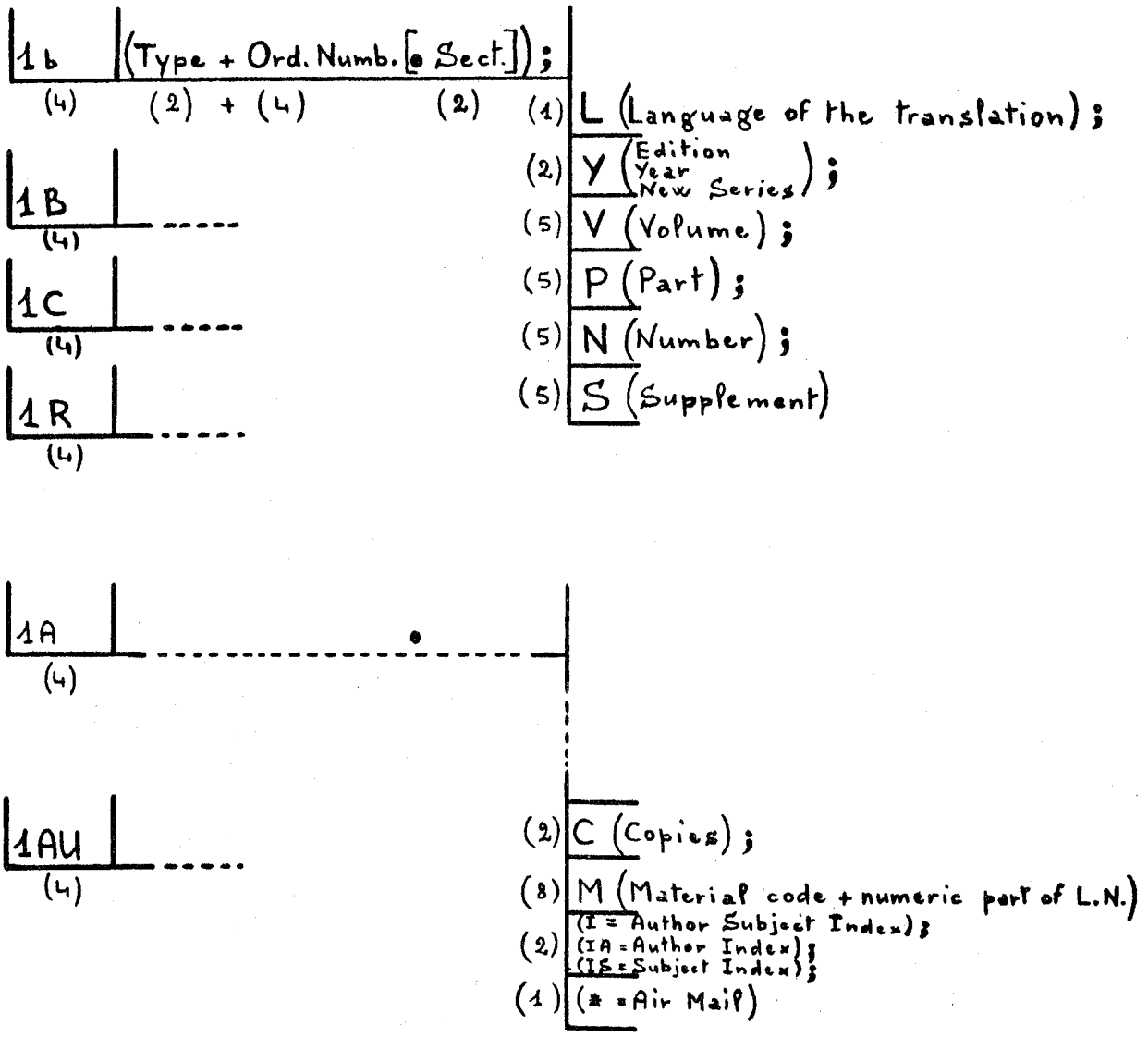

1b
(Type + Ord. Numb. [• Sect.]);
(4)
(2) + (4)
(2)
1B
(4)
1C
(4)
1R
(4)
(1) L (Language of the translation);
(2) Y (Edition Year New Series);
(5) V (Volume);
(5) P (Part);
(5) N (Number);
(5) S (Supplement)
1A
(4)
1AU
(4)
(2) C (Copies);
(8) M (Material code + numeric part of L.N.)
(2) (I = Author Subject Index);
(IA = Author Index);
(IS = Subject Index);
(1) (* = Air Mail)

fig. 2

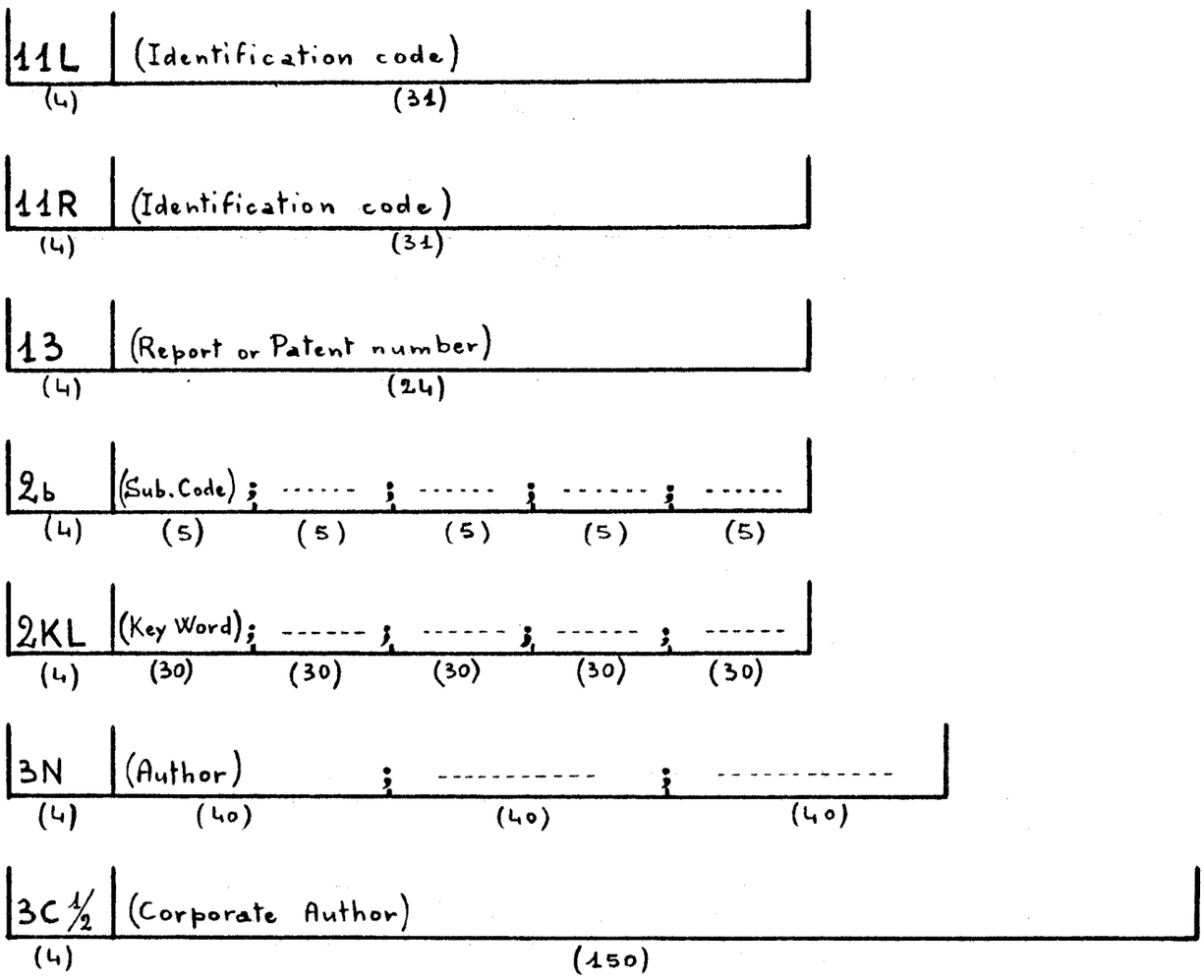

11L
(Identification code)
(4)
(31)
11R
(Identification code)
(4)
(31)
13
(Report or Patent number)
(4)
(24)
2b
(Sub.Code) ; ----- ; ----- ; ----- ; -----
(4)
(5) (5) (5) (5) (5)
2KL
(Key Word); ----- ; ----- ; ----- ; -----
(4)
(30) (30) (30) (30) (30)
3N
(Author) ; ----------- ; -----------
(4)
(40) (40) (40)
3C 1/2
(Corporate Author)
(4)
(150)

fig. 3

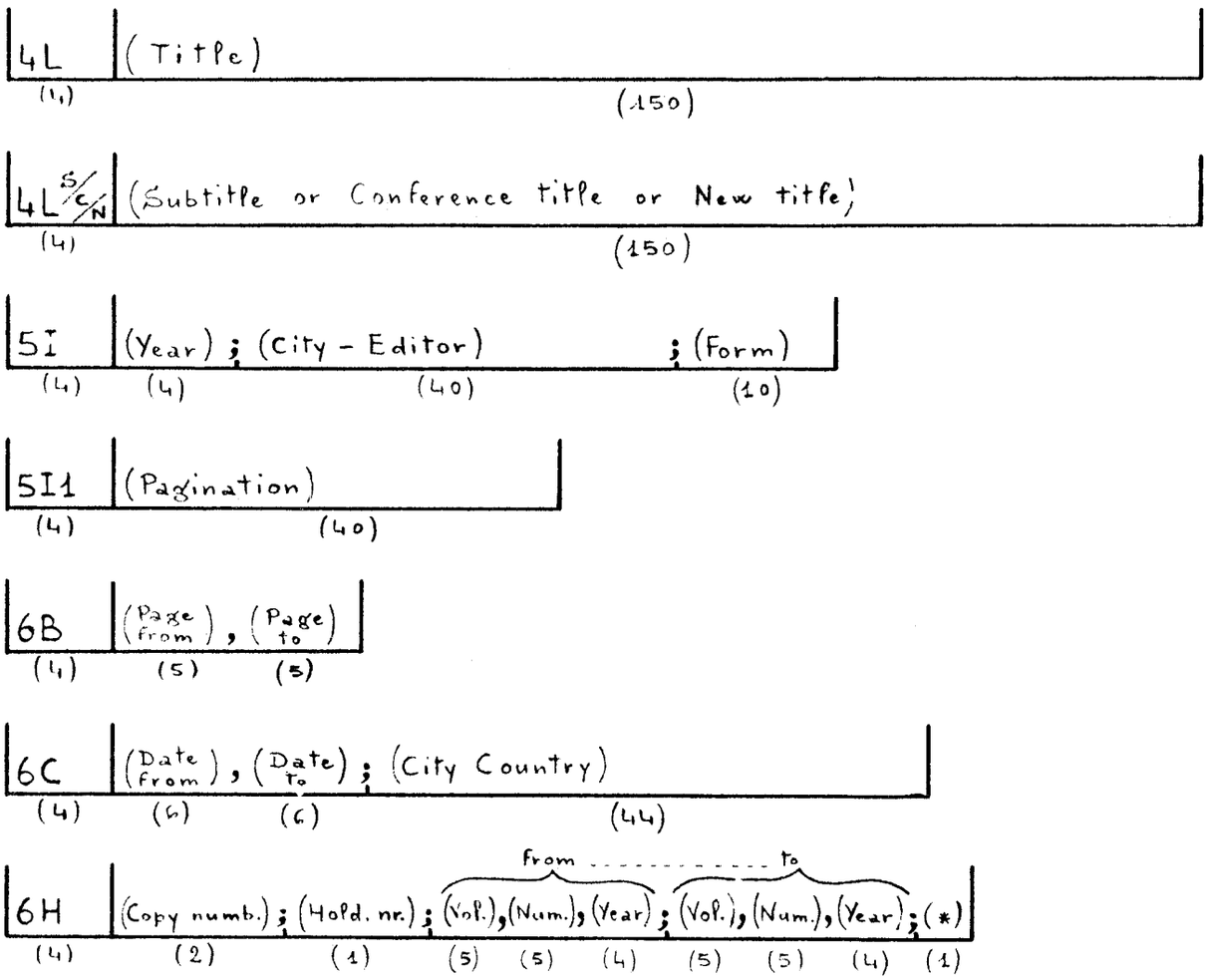

4L
(Title)
(1)
(150)
4L S/C/N
(Subtitle or Conference title or New title)
(4)
(150)
5I
(Year) ; (City - Editor) ; (Form)
(4) (4) (40) (10)
5I1
(Pagination)
(4) (40)
6B
(Page from) , (Page to)
(4) (5) (5)
6C
(Date from) , (Date to) ; (City Country)
(4) (6) (6) (44)
6H
(Copy numb.) ; (Hold. nr.) ; from (Vol.), (Num.), (Year) ; to (Vol.), (Num.), (Year) ; (*)
(4) (2) (1) (5) (5) (4) (5) (5) (4) (1)

Fig. 4

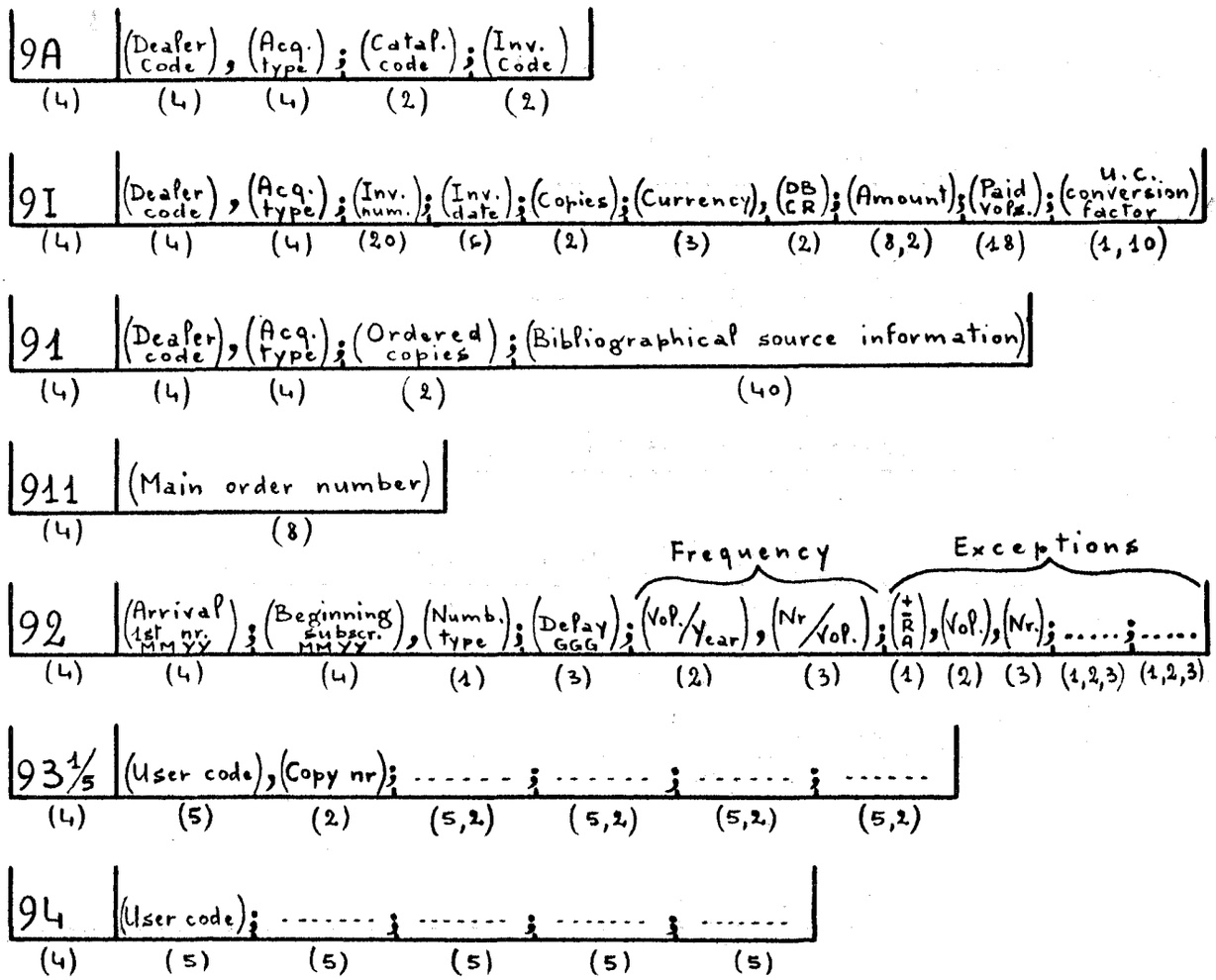

9A
(Dealer Code), (Acq. type); (Catal. code); (Inv. code)
(4) (4) (4) (2) (2)
9I
(Dealer code), (Acq. type); (Inv. num.); (Inv. date); (Copies); (Currency); (DB CR); (Amount); (Paid vol.); (U.C. conversion factor)
(4) (4) (4) (20) (6) (2) (3) (2) (8,2) (18) (1,10)
91
(Dealer code), (Acq. type); (Ordered copies); (Bibliographical source information)
(4) (4) (4) (2) (40)
911
(Main order number)
(4) (8)
Frequency
Exceptions
92
(Arrival 1st. nr. MMYY); (Beginning subscr. MMYY), (Numb. type); (Delay GGG); (Vol./Year), (Nr/Vol.); (+ R A), (Vol.), (Nr.); ....; ....
(4) (4) (4) (1) (3) (2) (3) (1) (2) (3) (1,2,3) (1,2,3)
93 1/5
(User code), (Copy nr); ------; ------; ------; ------
(4) (5) (2) (5,2) (5,2) (5,2) (5,2)
94
(User code); -------; -------; -------; -------
(4) (5) (5) (5) (5) (5)

Fig. 5

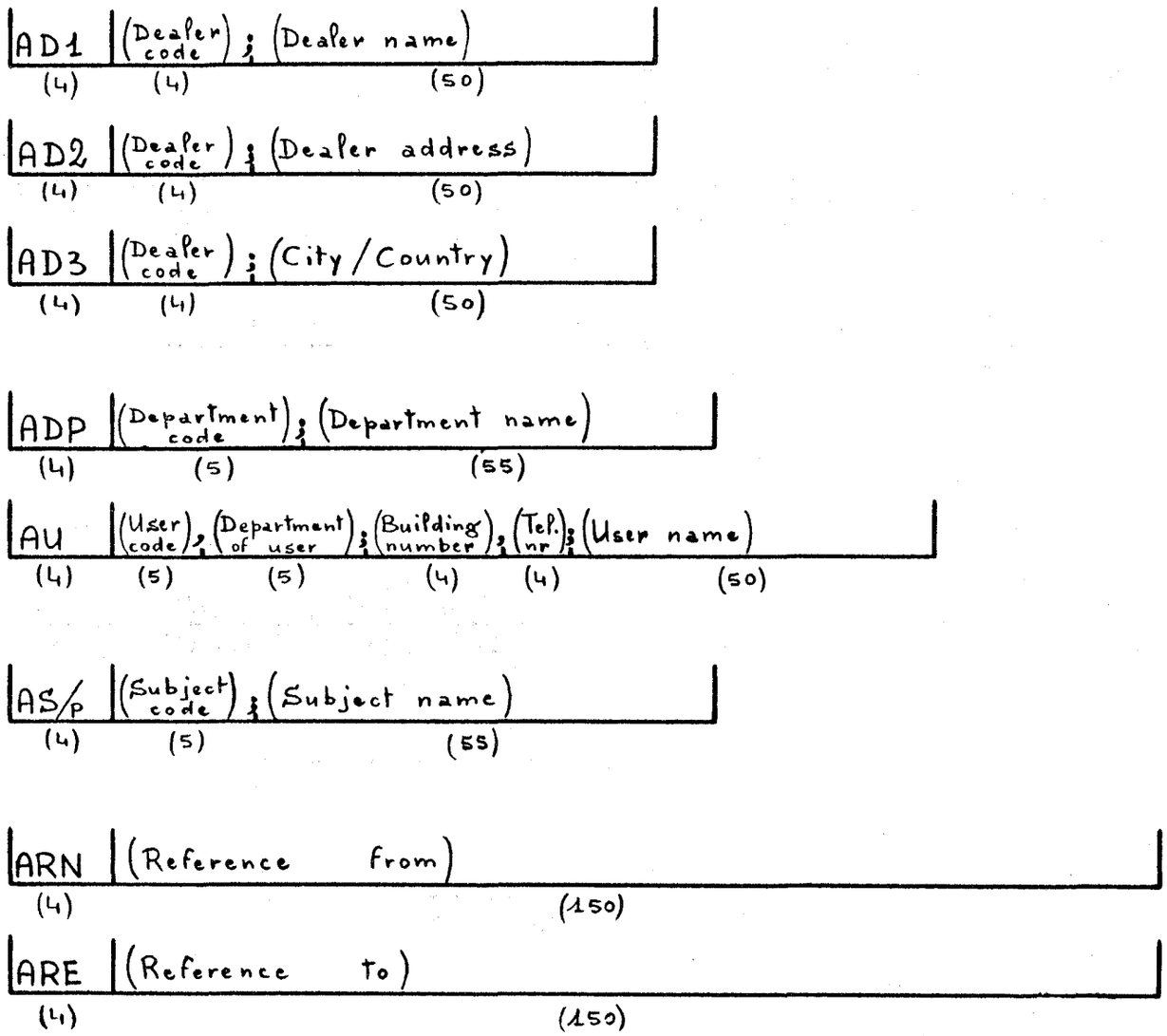

AD1 (Dealer code) ; (Dealer name)
(4) (4) (50)
AD2 (Dealer code) ; (Dealer address)
(4) (4) (50)
AD3 (Dealer code) ; (City / Country)
(4) (4) (50)
ADP (Department code) ; (Department name)
(4) (5) (55)
AU (User code) , (Department of user) ; (Building number) , (Tel. nr) ; (User name)
(4) (5) (5) (4) (4) (50)
AS/P (Subject code) ; (Subject name)
(4) (5) (55)
ARN (Reference from)
(4) (150)
ARE (Reference to)
(4) (150)

Fig. 6

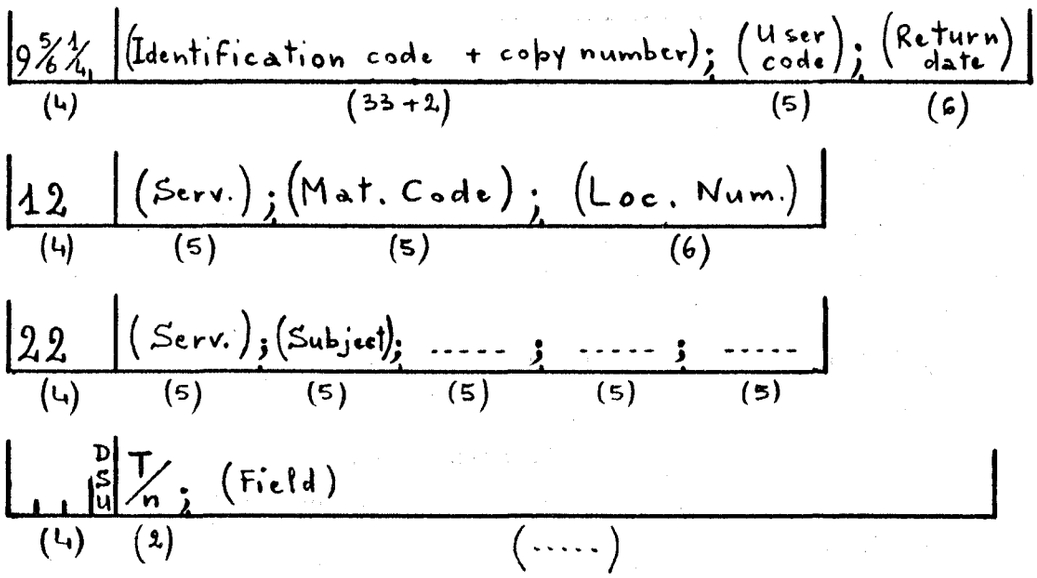

9 5/6 1/4
(Identification code + copy number); (User code); (Return date)
(4)
(33+2)
(5)
(6)
12
(Serv.); (Mat. Code); (Loc. Num.)
(4)
(5)
(5)
(6)
22
(Serv.); (Subject); ----- ; ----- ; -----
(4)
(5)
(5)
(5)
(5)
(5)
DSU
T/n ; (Field)
(4)
(2)
(.....)

fig. 1

F i l e s

Mater File
Order File
Inventory File
Catalogues File
Loan File
Auxiliary Files : Dealer File
User File
Subject File
General Reference File
Statistics File
Miscellaneous File : Input and Error List
Order List
Order to Dealer
Claim to Dealer
Loan List
Claim to User
Other Special List

Fig. 8

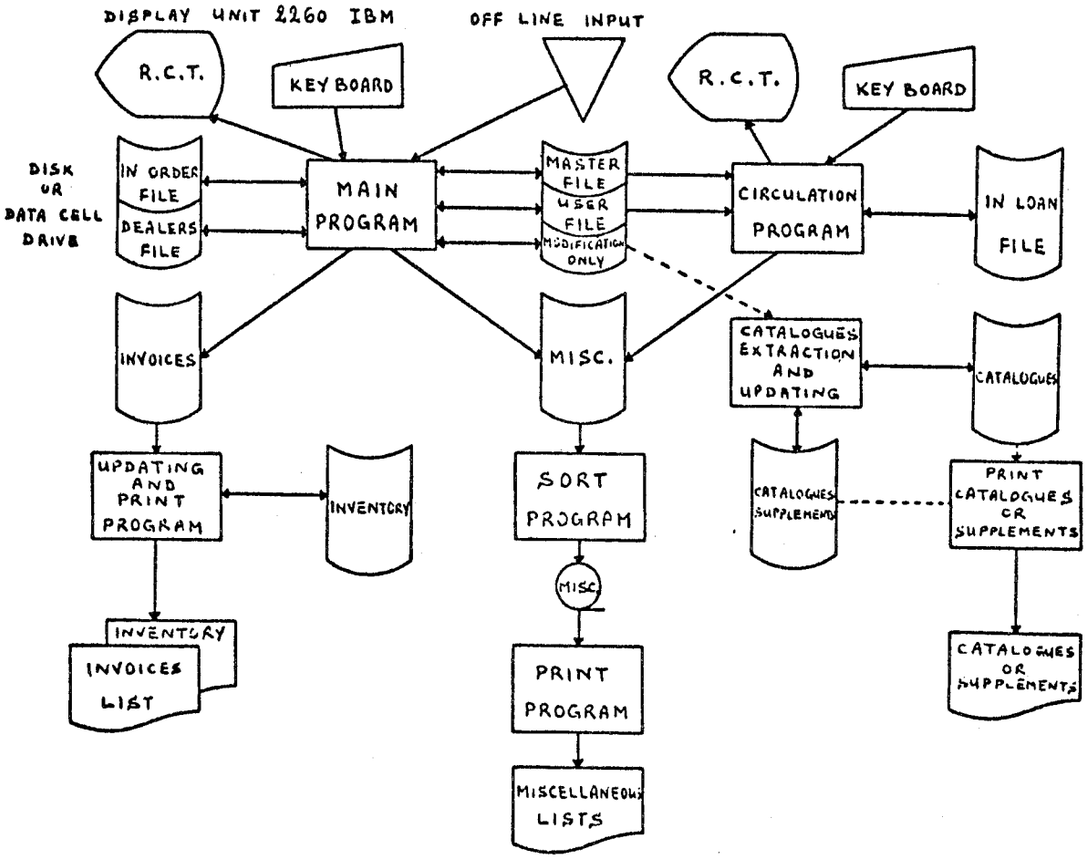

DISPLAY UNIT 2260 IBM
OFF LINE INPUT
R.C.T.
KEY BOARD
R.C.T.
KEY BOARD
DISK OR DATA CELL DRIVE
IN ORDER FILE
DEALERS FILE
MAIN PROGRAM
MASTER FILE
USER FILE
MODIFICATION ONLY
CIRCULATION PROGRAM
IN LOAN FILE
INVOICES
MISC.
CATALOGUES EXTRACTION AND UPDATING
CATALOGUES
UPDATING AND PRINT PROGRAM
INVENTORY
SORT PROGRAM
CATALOGUES SUPPLEMENTS
PRINT CATALOGUES OR SUPPLEMENTS
MISC
INVENTORY
INVOICES LIST
PRINT PROGRAM
CATALOGUES OR SUPPLEMENTS
MISCELLANEOUS LISTS

fig. 9

The Development of an Automatic Indexing System at EURATOM

G. LUSTIG

Abstract

The Scientific Information Processing Center (CETIS) of EURATOM ist developing an automatic indexing system. In view of future applications it aims at maximal compatibility with the manual indexing in the framework of the EURATOM Nuclear Documentation System.

Until now, an experimental system was realized by means of which a large variety of automatic indexing procedures was tested.

The paper describes the approaches examined, especially the use of statistical relations derived from a collection of abstracts and the corresponding manual keyword assignments. The results obtained are presented, and further possibilities of improvement are discussed.

1. Introduction (*)

In late 1966, the CETIS group "Automation Methods in Documentation and Translation" started a project with the principal objective of automating indexing within the framework of the EURATOM Nuclear Documentation System (1). For reasons of compatibility, the procedures to be developed must base on the following principles of manual indexing adopted in the existing system:

- English abstracts are used as source material.

- Each descriptor of the EURATOM Thesaurus (2) pertinent to a given abstract must be assigned. The thesaurus comprises 4665 descriptors.

- The relations contained in the glossary of the EURATOM Thesaurus must be respected. For each of the app. 10 000 glossary terms there exists either a "use" or a "see" relation. The "use" relation involves compulsory assignment of all the descriptors of the thesaurus referred to in the glossary. The relation "see" refers to one or more descriptors of which generally at least one should be assigned.

- In addition to the descriptors, the indexer may also assign any of the app. 6 600 "permitted" glossary terms and, if it seems to be necessary, any new (additional) term in order to increase the precision of indexing.

There is a very large number of possible solutions of automatic indexing which can only be evaluated experimentally.
Therefore, in the first phase of the project which will come to an end by the middle of this year, many of such procedures are tested. The best solutions found will be adopted for the design and realization of an automatic indexing system in the second phase of the project. This system should already be efficient enough for production purposes, but, on the other hand, a flexible structure should make it a useful tool for further experimental research.

This paper is to present some of the results obtained till now.

2. The test evaluation

The procedures to be tested are automatically simulated and evaluated by means of an experimental system which was realized at the beginning of the project. It consists of a series of IBM 360/65 programs and some IBM 7090 programs written earlier.

This way, for many indexing procedures the keyword assignments were obtained. The documents indexed were always the same : 105 abstracts chosen at

(*) This article is a modified and extended version of the paper presented at the meeting.

random out of the chapter "Physics" of Nuclear Science Abstracts, vol. 19 (1965), n° 11.

The evaluation of the tests is far from being perfect. The descriptors assigned manually to these documents within the framework of the EURATOM documentation system were used as term of comparison. This can be explained by the compatibility requested. The glossary and additional terms were disregarded in this comparison, since they are less regularly used in manual indexing than the descriptors. (Only 10 percent of the keywords assigned to the documents considered are glossary or additional terms). They will be taken into account when a more complete and reliable evaluation through retrieval tests will be possible.

Let be for a given automatic procedure
a the total number of descriptors assigned automatically, and
c the total number of descriptors assigned both automatically and manually.
With these numbers and the total number m of descriptors assigned manually, the consistency of two sets of keyword assignments for the same documents can be defined in different ways. Here, the quotient

$$q = \frac{c}{a + m - c}$$

was chosen, the denominator of which indicates the total number of assignments made in at least one of the two procedures compared. The "consistency factor" q varies between 0 (no common assignments) and 1 (all assignments are common). Its application is reasonable only if the two procedures yield almost the same "indexing depth", i.e. a and m must not differ too much from one another.

Often, procedures are tested which concern only certain special cases of assignments. Then, a is much smaller than m. For the discussion of such tests the "precision ratio"

$$\frac{a}{c}$$

is useful.

3. The pertinent concepts in an abstract

One of the main difficulties of automatic indexing is the determination of the pertinent concepts occurring in a given abstract. It will be discussed now in two steps.

The first one consists in determining all those expressions occurring in an abstract which might represent a concept of the scientific field considered. Generally, this is done by comparing the words of the abstract with the terms of a dictionary. The abstract words not found in the dictionary - if not yet known as non-significant words - can be evaluated separately for a future up-dating of the dictionary. The comparison of words of the abstract and the

terms of the dictionary depends on the criteria for their identification. They may depend, among others, on the following information :

- the number of the components of the dictionary term;

- the morphological form, in which the components are represented in the abstract. Here, three forms are distinguished :

  m_1 : complete identity, e.g. SPECTROMETER - SPECTROMETER

  m_2 : inflectional forms of the same words, e.g. SPECTROMETERS - SPECTROMETER

  m_3 : words with the same stem, e.g. SPECTROMETRY - SPECTROMETER

  m_2 includes m_1 and is included in m_3;

- the relative position of the abstract words put into relation to the components of the dictionary term, e.g. they may occur merely in the same sentence, or they have to appear in the same order without interruptions.

- the significance of the components. For example, the component METHOD of the term MONTE CARLO METHOD is so little significant that the occurrence of MONTE and CARLO can be taken as sufficient.

Each definition of the identification criteria leads to a different indexing procedure. Its efficiency essentially depends on the information contained in the dictionary.

The second step to be discussed here is the selection of the pertinent concepts out of all concepts identified in the abstract. For this purpose, frequency criteria can be used in automatic indexing of full texts. In an abstract, however, each word occurs only with a very low frequency which is practically irrelevant. Most approaches base on the questionable hypothesis that each concept occurring in an abstract is pertinent to it. The tables 1 and 2 are to illustrate to what extent manual indexing diverges from this hypothesis, but the consequences for the retrieval efficiency are not yet known.

The hypothesis is not used if the terms occurring in an abstract are interpreted as formal phenomena and related statistically to manual descriptor assignments performed previously. This method is described in chapter 5.

morphological form of the descriptors in the abstracts	u = number of descriptors in the abstracts	v = number of descriptors assigned man. as well as occurring in the abstracts	$\frac{v}{u}$	$\frac{v}{w}$
m_1	636	425	0.67	0.43
m_2	922	563	0.61	0.57
m_3	1454	638	0.44	0.65

(w = 981 = total number of non-compound descriptors assigned manually)

Table 1 : Occurrence and assignment of non-compound descriptors.

position of the components in the abstracts	morphologic. form of the components in the abstracts	u = number of descriptors in the abstracts	v = number of descriptors assigned manually as well as occurring in the abstracts	$\frac{v}{u}$	$\frac{v}{w}$
original uninterrupted sequence	m_1	51	38	0.75	0.09
	m_2	111	89	0.80	0.20
	m_3	114	90	0.79	0.21
arbitrary positions in the same sequence	m_1	74	46	0.62	0.11
	m_2	169	115	0.68	0.26
	m_3	198	123	0.62	0.28

(w = 438 = total number of compound descriptors assigned manually)

Table 2 : Occurrence and assignment of compound descriptors.

4. The representation of the pertinent concepts by descriptors

The tables 1 and 2 also confirm that the mere assignment of the descriptors occurring in an abstract is not sufficient. In fact, even assigning all descriptors the components of which occur in the morphological form m_3 and in any position within the same sentence (cf. the last lines of tables 1 and 2) one obtains for the 105 documents tested only c = 761 (= 54 %) of the m = 1419 descriptors assigned manually. Together with the large value of a - c (= 891) this makes the consistency factor small : q = 0.33. In the corresponding case in which only m_2 is permitted (cf. the last but one line of tables 1 and 2) the results are better, but not yet very good.

$$a = 1091, \quad c = 678, \quad q = 0.37$$

Given the condition of compatibility with the EURATOM documentation system, the most natural approach of a dictionary larger than the thesaurus is the glossary mentioned in chapter 1. The glossary relation "use" gives for the term at the left hand a unique representation by descriptors and is therefore well adapted for an automatic procedure. Instead, if there are two or more descriptors to which a "see" relation refers, there is an ambiguity which can only be resolved with some additional criteria.

Many indexing procedures defined and tested are basing on a dictionary comprising the glossary terms, the abbreviations of isotopes and chemical elements and the thesaurus descriptors - with references to their representation in the thesaurus. The best of these procedures identifies in a given abstract only those dictionary terms the components of which occur in the same sentence and in the morphological form m_2 . All relations are used without "see" references with more than two descriptors at the right hand. The procedure assigns a = 1645 descriptors ; c = 893 of them are also assigned manually. The consistency factor is q = 0.41.

This shows that the use of the glossary relations does not considerably increase the consistency factor (cf. q = 0.37 above). The difference would probably be larger in an evaluation basing on retrieval tests.

In principle, there are two possibilities of improving this dictionary. First, the glossary relations can be revised individually and better adapted for the automatic procedures. Also, the criteria for the identification of dictionary terms with words in the abstracts could be refined.

The second possibility of improvement is the extension of the relation system. More relations can be found in other thesauri, glossaries, tables, classification systems and indexes. Evidently, in such an extended system, the different relations are not necessarily to be applied in their original sense. Generally, this would create confusion and contradictions. Instead, each type of relation must be tested separately and interpreted

according to the test results. Further, the interlinkage and the superposition of different relations are to be investigated experimentally.

In the EURATOM project, the relation system is being extended by adding the references "see", "see also", "referred from by see", "referred from by see also" of the NSA Subject Heading System (3).

However, there have not yet been any test results. Another extension of the relation system will be discussed in the following chapter.

5. A statistical approach

A system of relations which can apply not only to pertinent but to all expressions of an abstract has been established by means of a statistical elaboration of abstracts and their manually assigned descriptors. Let be for any expression E of the abstracts and for any descriptor D of the thesaurus

$f(E)$ the number of abstracts considered in which the expression E occurs and

$h(E,D)$ the number of abstracts having the additional feature that D has been assigned manually.

The relative frequency

$$z(E,D) = \frac{h(E,D)}{f(E)}$$

approximates the probability for the correctness of the assignment of D to any abstract which contains the expression E. If $z(E,D)$ is higher than a certain level, one can define a relation between E and D and assign D if E occurs in an abstract.

This statistical approach leads, among others, to the following problems :

- It is necessary to fix how the occurrence of E in an abstract has to be defined. It is possible to distinguish different types of occurrence according to the criteria mentioned in chapter 3 of this paper. Then, $z(E,D)$ can be calculated and afterwards applied separately for each type of occurrence of E . Another way consists in introducing weights for the different types of occurrence, and in determining $f(E)$ and $h(E,D)$, to count each case of occurrence in accordance to its weight.

- If several expressions E of the abstract are statistically related to the same descriptor D, so, from their values $z(E,D)$ a new analogous value z must be calculated according to a formula to be defined. This way, also the attribution of weighting factors to the assignments could be introduced.

- Together with $z(E,D)$, other functions should be studied and tested (4). They may also depend on other variables than $h(E,D)$ and $f(E)$. A relation between E and D can be defined by means of several functions.

- In each collection of abstracts there are expressions occurring not often enough for a statistical evaluation. So, also this approach does not yield a complete relational system. In many cases, at least a less exact relation can be obtained if the frequencies are increased by means of less severe identification criteria.

In the CETIS project, only a first step towards a system of statistical relations was done until now. 741 abstracts out of the chapter "Physics" of NSA, vol. 19, no. 11 which were not used in the indexing tests were evaluated. In the role of E, all significant words but no compound expression were considered. All words with the same stem were identified. Thus, the statistical elaboration was excessively crude, and only modest results could be expected. Table 4 gives some examples of the statistical relations obtained and Table 3 shows the test results of some indexing procedures using no other relations or dictionary entries. (For example, even descriptors occurring in an abstract are not always identified as being pertinent). One can see that the quotients c/a are rather far from the mean of the values z(E,D) of the statistical relations involved and even lower than the minimum of these values. This can be explained by the unreliability of the values z(E,D) ; i.e. the condition $h(E,D) \gg 4$ chosen with regard to the small size of the collection evaluated is not strict enough. However, the combination of the statistical method with the glossary approach led already now to a considerable improvement of the results.

definition of the statistical relations		a	c	$\frac{c}{a}$	$q = \frac{c}{a+m-c}$
h(E,D)	z(E,D)				
≥ 4	≥ 0.8	556	420	0.76	0.27
≥ 4	≥ 0.75	695	495	0.71	0.31
≥ 4	≥ 0.7	792	530	0.67	0.32
≥ 4	≥ 0.6	1147	675	0.59	0.36
≥ 4	≥ 0.5	1681	782	0.47	0.34

(m = 1419 = total number of descriptors assigned manually)

Table 3 : Results of some indexing procedures basing only on statistical relations.

abstract word E	descriptor D	f(E)	h(E,D)	z(E,D)
CALCULATED	NUMERICALS	203	108	0.53
CHARGES	INTERACTIONS	64	33	0.52
CHARGES	ELECTRIC CHARGES	64	34	0.53
CHARGES	CHARGED PARTICLES	64	26	0.41
ENERGY	ENERGY	259	157	0.61
INTERACTING	INTERACTIONS	126	119	0.94
MASS	NUMERICALS	66	30	0.45
MASS	MASS	66	54	0.82
MASS	INTERACTIONS	66	31	0.47
MOVING	INTERACTIONS	22	9	0.41
MOVING	MAGNETIC FIELD	22	11	0.50
MOVING	MOTION	22	14	0.64
MOVING	PLASMA	22	9	0.41
MOVING	FREQUENCY	22	9	0.41
POTENTIAL	INTERACTIONS	55	29	0.53
VELOCITY	ELECTRON	59	26	0.44
VELOCITY	PLASMA	59	28	0.47
VELOCITY	VELOCITY	59	54	0.91

Each abstract word E is standing for all other words with the same stem

Table 4 : Examples of statistical relations between abstract words and descriptors assigned manually.

6. Combination of the approaches

Because of the incompleteness of the relational systems used the two approaches were combined. The best procedures previously tested could be composed, simply by adding their descriptor assignments. This way, a procedure was defined by which a descriptor D is assigned to a given abstract if one of the following conditions is fulfilled :

- D is the name of an isotope or a chemical element, and its abbreviation occurs in the abstract

- D has at least two components which occur in the abstract in the morphological form m_2 (i.e. inflectional forms of the same word are identified) and in the same uninterrupted sequence.

- there is a relation "use" or "see" (without ambiguity, i.e. without "or") between D and a compound glossary term the components of which occur in the abstract in the form m_2 and in the same uninterrupted sequence.

- in the abstract there is a single word E for which the inequalities $h(E,D) \geq 4$ and $z(E,D) \geq 0.7$ hold.

To the 105 documents this procedure assigns a = 737 descriptors of which c = 582 (= 73 %) coincide with those assigned manually. Because of the low indexing depth the consistency factor is only q = 0.37. By many tests, it was found out that no looser condition - e.g. the occurrence of non-compound descriptors even in the morphological form m - can be used without restrictions in the definition of an optimal procedure. Such conditions must be strengthened by combining them with those from other procedures. For example, in the optimal procedure a non-compound descriptor D is assigned if the following two conditions are fulfilled simultaneously :

- in the abstract considered D occurs in the morphological form m_2

- there is in the abstract a word E - perhaps D itself - for which the inequalities $h(E,D) \geq 4$, $z(E,D) \geq 0.4$ hold.

This way, in the optimum the conditions of some 30 procedures are involved. It does not seem worthwhile to describe them in detail, for several of them have only a provisional character. They were introduced for compensating the rather poor basis of the statistical approach.

The optimal procedure yields the following results :

$$a = 1257 , \quad c = 900 , \quad q = 0.51$$

Exemples of its assignments are shown in figs. 1 - 3.

The consistency factor of the optimal procedure can be compared with that obtained by an inter-indexer test within the framework of the EURATOM documentation system. To the 105 documents used in the experiments, descriptors were assigned also by a second indexer group and the consistency factor q equals 0.49. The comparison indicates that

with the approaches adopted automatic indexing can reach the level of routine work in manual indexing. This conclusion coincides well with the experience acquired by other research groups (5), (6). The next steps of the CETIS project are to confirm the obtained results through tests with larger collections and to improve them by the extension of the dictionaries and a refinement of the statistical approach.

References

(1) Rolling, L. : A computer-aided information service for nuclear science and technology.
Journal of Documentation, 22 (1966) n°. 2, p. 93-115.

(2) EURATOM Center for Information and Documentation - CID :
EURATOM-Thesaurus. Indexing terms used within EURATOM's nuclear documentation system.
Report EUR 500.e (2nd edition)
Presses Académiques Européennes, Bruxelles. 1966.

(3) Upchurch, Robert E. (ed.) : Subject headings used by the USAEC Division of Technical Information.
TID-5001 (5th Rev.)
USAEC Division of Technical Information. July 1965.

(4) Fangmeyer, H. ; Lustig. G. : The EURATOM automatic indexing project.
Preprint of a paper to be presented at the IFIP CONGRESS 68, Edinburgh, August 5 - 10, 1968.

(5) Salton, G. : The SMART system - retrieval results and future plans.
in : Information Storage and Retrieval. Scientific Report ISR-11.
Department of Computer Science, Cornell University, Ithaca, New York.
June 1966.

(6) Stevens, M.E. : Progress and prospects in mechanized indexing.
Paper presented at the Symposium on Mechanized Abstracting and Indexing, Moscow, September 28 - October 1, 1966.

20762 THE EQUILIBRIUM DISTRIBUTION OF NUCLIDES IN MATTER AT HIGH TEMPERATURES. F. E. Clifford and R. J. Tayler (Department of Applied Mathematics and Theoretical Physics, Cambridge, Eng.). Mem. Roy. Astron. Soc., 69: 21-81(1965).

At temperatures between 10^9 and 10^{10}°K and at suitable densities, nuclear reactions may become so profuse that approximate statistical equilibrium is set up between the concentrations of the different nuclides. Current theories of stellar evolution suggest that this equilibrium can be set up in stars and the resulting e-process elements may be distributed through space by a stellar catastrophe such as a supernova explosion. Complete statistical equilibrium between all the types of particles which occur in a star can not be set up because the neutrinos, which occur in all beta decay reactions, have an effectively infinite mean free path. However, nuclear and electromagnetic interactions occur much more rapidly than beta decays and in a short time the nuclei and photons come almost into equilibrium while the total numbers of protons and neutrons in the system are essentially constant. Subsequently beta decay reactions gradually change the relative proportions of protons and neutrons and, although the neutrino losses prevent true equilibrium, it is possible that the beta decays and inverse decays come into balance and that the ratio of the total number of protons to the total number of neutrons tends to a limiting value. Calculations are reported for the whole range of values of the physical parameters which are thought to be important, at least in stellar interiors. Nuclidic abundances were found as a function of three parameters: the density, the temperature and the ratio of the total number of protons to the total number of neutrons. In all, results were obtained for 333 sets of values of the three parameters; and detailed nuclidic abundances, before and after beta decay, are tabulated for 96 of these cases. The accuracy of the calculation is limited by the uncertainties of nuclear data. A detailed discussion is given of the possible size of errors arising from these uncertainties. It is also pointed out that, if the properties of only a few nuclides are shown to be seriously in error, it is a relatively simple matter to modify the tabulated results so that they are still solutions of the equilibrium equations, but now refer to different values of the density and the ratio of total number of protons to total number of neutrons. (auth)

Manual Indexing	Automatic Indexing
ABUNDANCE	ABUNDANCE
BETA DECAY	BETA DECAY
CROSS SECTIONS	
DENSITY	DENSITY
	DISTRIBUTION
ELECTROMAGNETIC FIELDS	
ELEMENTS	
EMISSION	
ERRORS	
EXPLOSIONS	EXPLOSIONS
HIGH TEMPERATURE	HIGH TEMPERATURE
INTERACTIONS	
	ISOTOPES
LOSSES	
NEUTRINOS	NEUTRINOS
NEUTRONS	NEUTRONS
NUCLEAR REACTIONS	NUCLEAR REACTIONS
NUCLEI	
NUMERICALS	NUMERICALS
PHOTONS	
PROTONS	PROTONS
SPACE	
STARS	STARS
STATISTICS	STATISTICS
TABLES	TABLES
TEMPERATURE	
THERMODYNAMICS	
THERMONUCLEAR REACTIONS	

Fig. 1 : Abstract no. 20762 from NSA, vol. 19, no. 11 with its manual and automatic descriptor assignments.

20816 EFFECTS OF MULTIPLE STIMULATION IN THE INTERACTION OF SEVERAL WAVES OF BOSONS. Guy Mayer (Compagnie Generale de T.S.F., Orsay, France). Compt. Rend., 260: 477-80(Jan. 11, 1965). (In French)

Elementary interactions between light and matter, where the number of bosons produced can be higher by several units than the number of incident bosons, are considered. Such interactions, often too weak to be detected experimentally in normal conditions can produce considerable phenomena whose excitation is intense enough to cause stimulation effects. The case of a photon producing, by interaction with matter, several identical phonons and one photon is discussed particularly because of its possible incidences on stimulated Raman effects and on the development of stimulated systems with increasing gain. (tr-auth)

Abstract

Manual Indexing	Automatic Indexing
BOSONS	BOSONS
ELEMENTARY PARTICLES	
ENERGY LEVELS	
	EXCITATION
INTERACTIONS	INTERACTIONS
LIGHT	LIGHT
PHOTONS	PHOTONS
QUANTUM MECHANICS	
SCATTERING	SCATTERING
SOLIDS	

Fig. 2

Abstract no. 20816 from NSA, vol. 19, no. 11 with its manual and automatic descriptor assignments.

21246 (TID-21697) COMPUTER PROGRAM FOR NUCLEAR MAGNETIC AND QUADRUPOLE RESONANCE FREQUENCIES AND INTENSITIES, I = ½ TO 9. Edward H. Carlson (Alabama Univ., University. Dept. of Physics). Feb. 1, 1965. Contract AT(40-1)-3090. 14p. Dep.(mn); $1.00(cy), 1(mn) CFSTI.

Two computer programs are described: HERDAG gives the eigenvalues and eigenvectors of a complex Hermitian matrix of any size; and ZEEMAN uses HERDAG to find the energy level schemes of any nuclear system having I = ½ to 9 in a magnetic field that is arbitrarily oriented with respect to the principal axis of the quadrupole interaction tensor. ZEEMAN also calculates the frequencies and relative intensities of transitions, in the case in which an arbitrarily oriented oscillating r-f field is present. (T.F.H.)

Abstract

Manual Indexing	Automatic Indexing
COMPUTERS	COMPUTERS
CONFIGURATION	
ENERGY LEVELS	ENERGY LEVELS
FREQUENCY	FREQUENCY
INTERACTIONS	INTERACTIONS
MAGNETIC FIELDS	MAGNETIC FIELDS
MATHEMATICS	MATHEMATICS
MATRICES	
NUCLEAR MAGNETIC RESONANCE	NUCLEAR MAGNETIC RESONANCE
NUCLEI	
	OSCILLATIONS
PROGRAMMING	PROGRAMMING
RADIO WAVES	
SPIN	
TRANSIENTS	
VECTORS	VECTORS

Fig. 3

Abstract no. 21246 from NSA, vol. 19, no. 11, with its manual and automatic descriptor assignments.

The Use of Machine Translation in Documentation

S. PERSCHKE

Abstract

The applications of the Russian-English MT system at CETIS as an instrument for information and documentation are presented. Four principal points are discussed:

- the Russian-English MT service at the request of the customers;
- current awareness with the automatic translation of the tables of contents of Russian periodicals;
- SDI with automatically translated abstracts from Russian periodicals;
- automatic indexing of Russian abstracts to be used as input for an IR system or as a key for SDI with user profiles.

At the end, the new Russian-English MT system which ist at present being implemented at CETIS is presented.

1. Introduction

Although MT, from the linguistic point of view is to be considered just at the beginning of the development, the translation quality presently obtained is sufficiently high for practical use.

CETIS is the only institution in Europe and one of three all over the world to provide a Russian-English machine translation service. The MT system in operation at CETIS, originally had been developed by the Institute for Languages and Linguistics of the Georgetown University, Washington (1). Its primary version became available to Euratom through a research contract in 1963, and it has since been steadily enlarged and improved. The translation system, at present, works at a speed of app. 60,000 Russian words translated per hour, and the over-all cost of some seven dollars per 1,000 words is highly competitive with that of human translation.

2. Applications

The practical applications of the MT system at CETIS are realized on different levels and range from a Russian-English translation service to advanced documentary applications which should operate on an automatic information system. In the following, the principal applications are discussed.

2.1. The MT service

To provide for translations of documents written in a language unknown to the user is a basic function in an information system. This function is fulfilled at Ispra, as far

as Russian is concerned, by means of the MT system (2).

Russian texts are translated automatically at the request of the investigators at Ispra and other Euratom Centres. Although no systematic publicity has been made for this facility, request has been steadily growing since its introduction. Statistics on the number of texts translated are available since 1965 :

1965	app.	30
1966	"	60
1967	"	120
1968 (to April, 1st)	"	30

The 120 documents translated in 1967 correspond to some 700,000 words.

The MT output is delivered to the customers, without any manual editing, together with a copy of the Russian text so as to enable them to identify formulae, equations, graphics etc. which could not be reproduced by the printer of the computer. The MT samples given in Figs. 1-4 may give an idea of the actual MT quality. Although it is all but perfect, experience has proved that, in general, it is sufficient to fulfill its primary function, i.e. to convey the information contained in a document to the customer. Although we did not perform any systematic evaluation of MT quality, a good indication for its acceptability is the fact that the customers normally are satisfied, and never have recurred to the alternate possibility of making retranslate the same text by man, were they dissatisfied with the machine output.

At present, we are examining the possibility of using MT as an aid to translators, so as to increase their efficiency and to produce more, and, possibly, better, translations. It should be possible to produce an equivalent to a scientific translator through post-editing the machine output by English language staff without knowledge of Russian, if there is the possibility of consulting a bilingual translator in the case of doubts.

HIGH TEMPERATURE SENSING ELEMENTS OF STRAIN GAUGE ON THE BASIS OF HEAT RESISTANT OXIDES

Mechanical Engeneering No 2 , 1967 G. UDC 536.453

L. S. Il'inskaya (Moscow)

The measurement of static deformations upon high (more 500 deg.) temperatures represents , as is known , important technical problem , which did not obtain up to the right time of satisfactory solution .

Up to the recent time to the creation of sensing elements of strain gauge , reliably operating in the conditions of high temperatures , hindered the absence of ribbon , which possesses by necessary properties . In recent years in the institute of precision alloys tsniichm worked out alloys and obtained ribbon , which with known limitations can be used for sensing elements of strain gauge , operating at temperatures 600 - 800 deg. . The successful forms of strain gauge ribbon were obtained also for boundary .

Being used in strain gauge as adhesive materials different cements (such , as vn-15 t , in-58 etc.) , well operating up to 500 deg. , possess by low electro-insulating properties and bad adhesion to metals at the temperatures above 500 of deg. .

Figure 1
Machine translation of a Russian article.

2.2. Current awareness of Russian literature

One of the difficulties of the access to Russian publications (as well as for other little known languages) is the fact that also the titles are uncomprehensible to the customer. Only part of the Soviet periodicals publishes tables of contents in English. For the rest, the investigator is bound to wait for references in secondary literature as NSA, or citations in other reviews.

To facilitate access to Soviet publications, therefore, in the environment of the Ispra Centre, the tables of contents are translated automatically and diffused in the Centre with the internal publication "NEW TITLES" which appears more or less weekly (Fig. 2 is to illustrate the presentation of the titles translated). This service was introduced at the end of 1966, and, as far as the MT service is concerned with, it had a double effect : on one hand, the demand for translations was doubled within 1967 ; on the other hand, the time lag between publication and access (i.e. request of translation) was considerably reduced. This can be seen from the following table :

time lag between publication and translation

years	less than 6 months	6 - 12 months	more than 12 months
1966	10 %	20 %	70 %
1967	30 %	30 %	40 %

2.3. SDI of Russian abstracts

The titles, as it is well known, contain insufficient information about the usefulness of an article. Therefore, we are at present examining the modalities of enlarging the current awareness service with some kind of SDI with automatically translated abstracts of Russian periodicals. (Fig. 3 is a sample of such an abstract).

ZAVODSKAYA LABORATORIYA

INDUSTRIAL LABORATORY

VOL.33 (1967) NO5

Content

The Methods Of Chemical Analysis

Figure 2
Machine translation of a table of contents.

UDC 681.17.001.5 .

The Method Of Calculation Of Dynamic Characteristics Of Measuring Systems , Which Include Manometers And DIFMANOMETRY .

Preobrazhenski V. P. , Ivanova G. M. .

" TEPLOENERGETIKA " , 1968 G. , N/o 2 , 72 .

Are presented the results of investigation of dynamic characteristics of measuring systems , which include initial apparatuses with the impulse lines and secondary electronic apparatuses .

Given in article the method of calculation of range of working frequencies and the dynamic characteristics of measuring systems necessary upon the appraisal of the dynamic errors of the latter and for the regular selection of apparatuses upon the recording of the nonstationary processes .

Tables 1 .

Illustrations 3 .

Bibliographies 5 .

Figure 3
Machine translation of a Russian abstract.

It is known that abstract journals as NSA, for soviet publications, have a time lag of at least 6 months. Therefore, a properly organized MT service for abstracts could very well close a gap in the existent information systems. We did not make any decision about the modalities and distribution of such a service, but we feel that it might be useful even for a larger community than the Euratom Joint Research Center.

2.4. Automatic Indexing of Russian Abstracts

Last year, we carried out an experiment of assigning automatically English keywords of the EURATOM Thesaurus to original Russian abstracts (3). The experiment was rather limited - it comprised a collection of some 70 abstracts from the field of plasma physics and astrophysics which were also referenced by the NSA and indexed manually within the framework of the EURATOM Nuclear Documentation System (4). The analysis of the documents produced some 500 indexing terms which were integrated into the dictionary. The experiment proved that, from the technical point of view, bilingual indexing does not present particular difficulties. The indexing procedure itself which had been adopted - mainly the formal match of text words with indexing terms and the application of glossary relations (5) - was rather brute-force, however, we believe, it should be improvable, especially in connection with the automatic indexing project of CETIS which is basically language-independent (6).

The principal advantage of such an application is, again, the timeliness and also the economy. If the abstracts are also translated - which is highly desirable - indexing is a by-product at practically no extra cost.

Fig. 4 is to illustrate the output of an abstract translated and indexed automatically. Fig. 5 reproduces the same document abstracted in NSA and indexed by CID. As one can see, a direct comparison between the two samples is not possible, because the document in NSA is not a mere translation of the Russian abstract written by the author, but much more detailed. The

5 V290

CONCERNING ONE POSSIBILITY OF INVESTIGATION OF COMPOSITION OF PRIMARY COSMIC RADIATION OF ULTRAHIGH ENERGY

NESTEROV N. M. , NIKOL6SKI S. I.

((BULL. ACAD. SCI. USSR .
SER. PHYS.)) , 1964 , 28 , N/O 12 , 1930 - 1933

WAS EXAMINED THE POSSIBILITY OF INVESTIGATION
OF COMPOSITION PRIMARY COSMIC THE RADIATIONS OF ULTRAHIGH
ENERGY ACCORDING TO FLUCTUATIONS REL. THE INTENSITIES
OF CHERENKOV FLARE OF LIGHT UPON THE PASSAGE OF WIDE SHOWER
THROUGH ATMOSPHERE .
THERE IS CONDUCTED THE COMPARISON EXPERIMENT THE DATA WITH
COMPUTATIONS , WHICH WERE MADE UPON DIFFERENT PREMISES
CONCERNING THE COMPOSITION OF PRIMARY PARTICLES .
ANALYSIS SHOWS , THAT COMPOSITION PRIMARY COSMIC RADIATIONS
WITH ENERGY EV , ACCORDING TO-VISIBLE , DOES NOT
DIFFER FROM THE COMPOSITION OF PRIMARY RADIATION IN
THE RANGE OF ENERGIES EV .

KEYWORDS ASSIGNED TO THE ABOVE DOCUMENT

PRIMARY COSMIC RADIATION
COSMIC RADIATION
RADIATIONS
ENERGY
EXTENSIVE AIR SHOWERS
COSMIC SHOWERS
ENERGY RANGE
SHOWERS
ATMOSPHERE
MEASUREMENT
NUMERICALS
PARTICLES
ANALYSIS
EV RANGE

5 В290. **Об одной возможности исследования состава первичного космического излучения сверхвысокой энергии.** Нестерова Н. М., Никольский С. И. «Изв. АН СССР. Сер. физ.», 1964, 28, № 12, 1930—1933

Рассмотрена возможность исследования состава первичного космич. излучения сверхвысокой энергии по флуктуациям относит. интенсивности черенковской вспышки света при прохождении широкого ливня через атмосферу. Проводится сопоставление эксперим. данных с расчетами, сделанными при различных предположениях о составе первичных частиц. Анализ показывает, что состав первичного космич. излучения с энергией $\sim 10^{15}$ *эв*, по-видимому, не отличается от состава первичного излучения в интервале энергий 10^{10}—10^{12} *эв*.

Figure 4
Machine translation and automatic indexing of a Russian abstract. The NSA abstract and the manual indexing of the same document are shown in Figure 5.

20883 A POSSIBILITY OF INVESTIGATING THE COMPOSITION OF THE PRIMARY COSMIC RADIATION OF SUPERHIGH ENERGY. N. M. Nesterova and S. I. Nikol'skii (Inst. of Physics, Academy of Sciences, USSR). Izv. Akad. Nauk SSSR, Ser. Fiz., 28: 1930-3(Dec. 1964). (In Russian)

An analysis of the composition of primary cosmic rays of more than 10^{14} ev was made, based on Cherenkov flashes occurring when a large cosmic shower was passing through the atmosphere and on the number of particles at the observational level. The fluctuations of the ratio of Cherenkov flashes Q versus the number of particles in the shower n obtained at Pamir (elevation 3860 m) were compared with the computations of the composition of primary cosmic radiation based on various assumptions of its components. Using the ratio Q/n, the composition of primary cosmic rays in showers was computed for two assumed types of protons and other heavy particles. The first type contained data about the composition of primary cosmic rays with energies of $10^{10}-10^{12}$ ev at the upper limit of the atmosphere. The second type contained primary cosmic rays with a composition having heavy nuclei with particle energies from 10^{11} to 10^{15} ev. The distribution of particles depends upon the composition of the primary cosmic rays. (ATD)

ABUNDANCE
ANALYSIS
ATMOSPHERE
CHERENKOV RADIATION
COSMIC RADIATION
ENERGY RANGE
EXTENSIVE AIR SHOWERS
LEVELS
MEASUREMENT
NUCLEI
NUMERICALS
PAMIR
PROTONS
SCATTERING
SHOWERS

Figure 5
Reproduction of the abstract 20883 of NSA, Vol. 19, No. 11 and its manual indexing.

automatic assignment of index terms to the Russian abstracts translated should be useful not only in an IR system, but also for the development of a fully automatic SDI system.

3. MT Development at CETIS

In order to increase the efficiency of its machine translation service, since 1963 CETIS has performed the following improvements of the translation system :

- a detailed analysis and description of the computer programs (7), (8), which went along with the re-programming of the system under the control of the IBM 7090 IBJOB monitor system as to reduce considerably operator interventions and to increase the performance of the program ;
- periodical updating of the dictionary and improvement of some linguistic operations which increased somewhat the translation quality ;
- a modification of the input conventions as to permit a more efficient control of processing non-Russian items in the source text, especially, in order to avoid nonsense matches with Russian dictionary entries ;
- an enlargement of the input media as to increase the input capacity. It is now possible to keypunch Russian texts not only on punched cards, but also on paper tape with either Russian or English key-board ;
- the introduction of an output with upper and lower case characters, which highly increases the legibility of the translations and eliminates a certain psychological resistance to the characteristic all-upper case machine output. (Compare the samples given in Figs. 1-3 with the older one in Fig. 4). Since spring 1968, all MT output has been printed with the new facility.

These modifications are very useful for our production purposes, but they do not or only marginally influence the translation quality itself. More important linguistic improvements have

not been achieved, the rather poor basis of the actual system making them practically impossible. One should not forget that the Georgetown University system was the first one in the world to be started, and certainly suffered from the unlimited optimism of the pioneer period in which MT was considered basically as a one-to-one term substitution with the addition of a few rules concerning the differences between the source and the target language (9). This concept very soon turned out to be inadequate, but the Georgetown University project never revised it completely. Thus, actually the entire set of linguistic operations in the system is a long series of frequently contradictory ad-hoc solutions, and it is practically impossible to predict the effect of modifications or additions to the analysis performed.

Therefore, in order to obtain a sensible improvement of the translation quality, CETIS is developing now a new system (10). Its design bases principally on the experience and the criticism of the Georgetown system. The main objectives of the new system are :

- a new design of the algorithms of linguistic analysis, especially of the syntax, in a way that one exploits first the formal information already contained in the dictionary and in a second phase adds gradually new information, principally of semantic nature. The primary purposes of the design is to make the system open-ended ;
- the integration of a larger dictionary (180,000 entries against 30,000 in the actual system). The strategy of dictionary look-up presently adopted does not permit the use of such a large dictionary. The new strategy bases on special list- processing techniques ;
- a new design of the special-purpose programming language SLC (7), (8), as to make it basically computer-independent and more flexible, in particular for other linguistic applications ;
- an optimal exploitation of the resources of the IBM 360/65 installed at CETIS. This will raise the translation speed to app. 300,000 words/hour.

The synthesis of these objectives permits various improvements of the actual translation procedure as e.g.

- the handling of non-Russian items. While in the Georgetown system, Russian and non-Russian items (as figures, formulae, English words, etc.) are looked-up indifferently in the dictionary (occupying time and space) and can produce accidental nonsense matches (Cu translated as Chew), in the new system they are considered as a mere character string and identified linguistically through a code which is attached to every non-Russian item ;
- the handling of compound words. While with a dictionary recorded on magnetic tape the only economically acceptable solution is the detaching of certain pre-defined prefixes (as pseudo- semi - poly- etc.) during the text input, the use of a disk storage permits very well to analyse compound words during the dictionary search. Thus, free word combinations which in fact are unpredictable and cannot possibly be contained all in the dictionary as "JELEZOXROMOALHMINIEVYl" (iron-chromium-aluminum) can be identified and translated ;
- the handling of homographs - i.e. multiple matches of text words with dictionary entries. While actually, the problem is disregarded, except some accidental cases (e.g. TOM), it is provided to include all possible matches into the translation process, and to reduce the number of alternatives - possibly to one - in the course of syntactic analysis. Homographs which cannot be resolved with the general analysis procedure are treated individually.

Also, the field of documentation will profit from the new system. The improvement of the translation quality and economy will permit a better exploitation of the existing possibilities of application, while the new SLC system, because of its flexibility, will become an appropriate tool for more advanced documentation problems involving linguistic features.

References

(1) Dostert, Leon E. : "The Georgetown-I.B.M. experiment." in : Machine Translation of Languages. The Technology Press of M.I.T., Cambridge (Mass.). 1957.

(2) Perschke, S. ; Lustig, G. : Automatische Sprachübersetzung - Fünf Jahre praktischer Übersetzungsdienst Russisch-Englisch bei EURATOM. Atompraxis, (1968) Heft 4/5.

(3) Perschke, S. : The use of the "SLC" system in automatic indexing. Proceedings of the F.I.D./I.F.I.P. Conference 1967, North Holland Publishing Company, Amsterdam.

(4) Rolling, L. : A computer-aided information service. Journal of Documentation, 22 (1966) No. 2, p. 93-116

(5) EURATOM-Center for Information and Documentation - CID : EURATOM-Thesaurus. Keywords used within EURATOM's nuclear documentation system. Report EUR 500.e Presses Académiques Européennes, Brussels, 1964/1966

(6) Lustig, G. : The development of an automatic indexing system at EURATOM. These Proceedings

(7) Brown, A.F.R. : The "SLC" programming language and system for machine translation. Report EUR 2418.e. Presses Académiques Européennes, Brussels. 1965

(8) Perschke, S. : The computer programs of the "SLC" system for machine translation. Report EUR 2583.e. Presses Académiques Européennes, Brussels, 1965

(9) Perschke, S. : Automatic language translation, its possibilities and limitations. EURATOM Bulletin, (1967) No. 2

(10) Perschke, S. : Machine translation, the second phase of development. Endeavour, 27 (May 1968) No. 101, p. 97-100

Contacts Between the Euratom Nuclear Documentation System and its Users

by C. Vernimb

Abstract

Experience in operating the Euratom Nuclear Documentation System (ENDS) and the statistical evaluation of user comments yielded the following statements:

Average relevance ratio for documents retrieved by computer in retrospective literature searches: between 35 and 40 %. Checking the relevance of the documents before sending them to the user is necessary.

Average recall ratio for documents retrieved in retrospective literatire searches: round about 70 %.

Average processing time for each query: 2 weeks.

Satisfaction of user with the results: round about 90 %.

Average relevance ratio for documents retrieved in SDI searches: 75 %. Checking the relevance of documents before sending them to the user is not necessary due to better query formulation.

Average recall ratio for documents retrieved in SDI searches: more than 70 %.

The average relevance ratio for documents retrieved in retrospective literatire searches is expected to increase from now about 40 % to more than 80 % by a sample screening and relevance feedback method resulting in a considerably reduced processing time.

"The proof of the pudding is in eating". These were always Mr. Brée's words when, during the last few years, people from inside and outside the CID expressed their opinion that the Euratom Nuclear Documentation System would perhaps not work well, due to this or that reason.

I am now going to give you some facts about "the eating". In other words: Which are the customers reactions on our sercice.

RETROSPECTIVE SEARCHES

Up to now we processed close to 1,500 queries. If you are of those who already sent us queries and received the results you will know that we ask from all customers a detailed appreciation of the material supplied: As long as we do not ask for another payment of the literature searches, this "feedback" is the only payment required.

In the feedback the customer is expected to tell us which of the supplied references he knew already, which are pertinent and were unknown to him, which are not just pertinent but useful, and which are not pertinent at all; finally the customer ist asked to indicate any pertinent references which we failed to send him.

By this procedure we get a fairly good picture of the customer's degree of satisfaction. Furthermore we do extensive checking as to why we did not retrieve the missing documents.

The following figures present some statistical data in retrieval and feedback.

Fig. 1 shows the number of queries received per month from February 1967 to April 1968 (full line). In addition are given the number of queries treated (dotted line) and the number of queries still to be treated at the end of each month (dash line).

The average number of queries received per month is 82. In May, June and July we made a great effort to reduce the number of queries still on hands, that means to deal as fast as possible with those queries we received. This led to short processing times of not much more than one week. However, during the holiday period it was not possible to keep this figure so low. The peak in the curve representing the number of queries received in October, November and December is due to an advertizing action undertaken in September.

Fig. 2 represents the number of abstracts screened, that means checked for relevance for the same period (full line), as well as the number of abstracts really sent out (dotted line). Both curves are in principle similar to those presented before, because the average number of abstracts screened per month and per query amounted to a hundred per month rather constantly.

The ratio of the two values for each month is considered as the average relevance ratio per month. This value os presented in Fig. 3. There is a clear increase in the average relevance ratio, now being round about 40 %. This mounting tendency is due to increased experience in query formulation.

Part of the references sent to the customers are designated as being relevant, the rest are designated as being "possibly relevant". For the possibly relevant items it was not feasible to make sure, on the basis of the abstract, whether the original document contained the information wanted. Another reason for sending possibly pertinent references is the fact that we did not always understand clearly the user's query. There is a tendency in favor of the pertinent references due to the fact that we more and more often ask, over the telephone, for more precise information (FIg. 4).

Fig. 5 gives the number of comments on the literature searches for each month for the same period (full line). The dotted line represents that part for which no further action was necessary because the customer was satisfied with the results and did not need further information. The dash line gives the number of comments for which further action was necessary.

Fig. 6 gives the idealized increase in the total number of queries received (full line) and the corresponding value for "comments" received (dotted line). From the gradient and the phase shift it can be deduced that 65% of the customers sent comments on queries, and that with an average delay of 3 months.

Fig. 7 represents the monthly averages of customers satisfaction, given in percentages of documents mentioned in all comments. The full line represents that part for which no further action was necessary.This curve corresponds to the dotted line in Fig.5.The dash line represents that part for which the customer agreed in principle with the results obtained, which we had sent as preliminary results,but asked for more information of the same kind.The total of these two percentages is represented by the dotted line.The part remaining between the dotted line and 100% represents those comments cases in which the customers stated that we did not meet their requirements, due to misinterpretation of their queries,or to changes in the customers delimitation of their own queries,and it can be seen from the dotted line that the total value for customer satisfaction increased to now round about 90%.

Fig. 8 gives the successive steps in query processing and the corresponding operation times.It is to be seen that considerable time is used for query formulation by the subject specialist,for punching and for machine processing of the query,for abstract card drawing,and especially for screening the abstracts, that is for separating relevant from non-relevant references. Fig.8 gives the pure operation times.But of the successive steps mentioned,an individual one can be performed only after completion of the preceding step, and therefore it can be calculated that the total time required for the processing of the query, even if the indispensable computer run is effected at night, will be about 24 hours.And this 24-hour-schedule can be achieved only if each step follows immediately upon the other. In practice there are many queries being handled in parallel, which fact will result in delay or waiting times.According to our experience the number of queries under treatment at any given moment has to be multiplied by 0.3,to get total processing time in days.So, if 40 queries have to be treated,their processing time is 12 days.

These statements are based on the number of staff members available now,that is an effective personnel as it has so far been disposable for input and system development,without the staff additions indispensable for current retrieval operations.

It is evident that certain functions must be maintained without any modifications, as for instance data input relating to new documents,control of keyword assignment, further development of the Thesaurus,evaluation of retrieval results and,last not least,training.In fact,during the actual experimental phase of the system,only about 30 to 40% of the working time of qualified staff members are available for current retrieval operations.

Another factor that will influence processing time will be the irregularity with which the queries come in.

Fig. 9 gives the number of queries received from day to day in November 1967. There are days on which no query came in. But there is also a day when 25 queries came in.

Fig. 10 gives a summary of retrieval and feedback evaluation: of a hundred abstracts screened for an average literature search, 62 are discarded as non-relevant (or "noise"), and 38 are sent to the user.

Of the 38 references 25 were indicated as pertinent, and 13 as possibly pertinent. Of those references as were indicated as pertinent, the user considered 5 as non-relevant and 20 as relevant, 5 of which he knew already. Of the 13 indicated as possibly pertinent, the user considered 6 as non-relevant and 7 as relevant, 2 of which he knew already.

From comparisons with manual searches very thoroughly done we know that there are, on an average, 13 pertinent references not retrieved by the semi-automatic system, one of which the user knew already.

Of the 13 missed references 8 are missed due to poor abstracting, more than 2 due to poor indexing, 2 due to poor retrieval, and less than 1 due to poor screening. This "less than one" is in fact included in the 62 discarded as non-relevant.

It has to be stated that the figures given are related to an average of several hundred queries treated neither with undue superficiality nor with excessive concern for exhaustiveness (the queries considered are real-life queries, not such especially selected for statistical purposes).

Fig. 11 gives percentage of documents fed into the computer memory, with respect to their subject field (white columns); the black columns represent the percentage of queries with respect to the same subject fields (basis of statistics: 1400 queries). There is an important difference in the "input" (more than 40 %) compared with the "output" (12 %) in Physics. On the other hand it should be noted that the

"output" in Metals and Ceramics, in Engineering and Equipment, and in Reactor Technology is 2 to 3 times as high as the "input". The reason for the relatively small number of queries in physics seems to be that physicists believe that they will not obtain new references, that means such as they do not know already, from our documentation system. In addition, physicists are often interested in properties of materials which are covered by other subject categories, whereas specialists from the other subject fields are not so often interested in documents on physics. In general it can be stated that there is a considerable interest in documents dealing with properties of materials, independently of specific subject categories. Another statement which can be made is that the "output" in industrial applications of isotopes is 6 times as high as the "input".

There are some more results gained by our experience on retrieval and by the evaluation of feedback:

1. The most difficult problem in communicating the query is its exact delimitation. In my opinion this is due to the fact that the average user is not accustomed with automated documentation systems.

   If a scientist has a problem and wants to get information, he first consults the books on his desk, or the books, journals and reports in the closest library. To start searching he needs a heading by which to enter the subject index of a book or the card catalog of the library. If he does not find an answer, or if he finds too many answers, he tries again under another entry. After checking a few of the references found, normally some other entries will come to his mind and he will start searching anew. This type of a manual literature search uses single term entries and applies what I would call "immediate feddback", whereas in a computerized documentation the searching is done by a man-machine-system using term coordination and delayed feedback.

In the first type of search one may start with only a vague idea of the query; the second type requires a precise query formulation, because it lacks the immediate feedback which adds precision to the query.

Our experience is that we receive vague queries (that means queries corresponding to single-term catalog entries) mainly from libraries. This is easy to explain: The originator of a query formulates his needs just as if he were starting to browse in the library; he passes it to the librarian who acts as a letter-box and sends it to us at the CID.

May I therefore ask you, Ladies and Gentlemen, the librarians to assist the scientists in precisely delimiting their needs, and help them to get accustomed with automated documentation systems.

At this point the question will arise whether we are not losing anything by switching from the manual system, with single-term entries and immediate feedback, to the machine system based on term coordination and delayed feedback? Is the problem only one of getting accustomed to a new system, or was the old manual system better suited to our way of thinking?

My opinion is the following: Obliging the originator of a query to express his needs more clearly is in any case to his profit: he will better understand his own problem. But if this is not possible, because the originator of the query does not know enough about his problem, the single-term-entry method with immediate feedback offers some advantages. Direct access to the Computer for every scientist is considered to be the solution in a later phase of development of our system.

2. In figure 12 the effort (in time or money) needed to obtain information is given for manual and machine searches:

In a manual search you can already gain some information with a small effort; for a machine search there is a threshold, a minimum effort required before the first bit of information is obtained. Therefore, for getting a relatively small amount of information, it does not pay to pass the query to a documentation center. For

such a preliminary query a library should be consulted, and it should be the task of the librarians to screen out such preliminary queries.

3. There is another relation expressed by the machine search curve: To retrieve 90 % of the available information the effort is about three times as high as for the retrieval of 60 % of the available information. According to the experience gained from our feedback campaign the customer will generally be satisfied with 60 % of the available information. Nevertheless most of the customers ask for "a complete bibliography". As long as we are not in a position to charge him different prices for a complete bibliography and for a general survey of the subject, we try to make it more difficult for the customer to ask for a complete bibliography by asking him to tell us how many documents on the subject he is aware of alreaedy, and how many he expects to receive. These questions are included in our new formsheets for requests.

4. Feedback does not only lead to a better understanding of the customer's queries, but also to a better understanding of the system itself. Most of the parameters of our system proved to have been quite reasonably chosen: Using abstracts as a basis for indexing instead of titles or full documents; using a controled vocabulary for indexing and retrieval, etc. If we had to establish a documentation system for another subject field we would not change these parameters. Perhaps would we use more specific terms and apply a more thorough dictionary control right from the beginning.

S D I

Retrospective searches are only one part of our output; the other part is SDI: Selective Dissemination of Information. Starting in January 1968, we processed batches of more than 10,000 references each month for an increasing number of customer profiles. We have now more than 100 profiles in operation and more than 50 in preparation.

SDI can be handled the way NASA or MEDLARS do: elaborate a profile and send the results without screening.

This would mean that we expect the user to accept 65 % of noise, according to our average relevancy of about 35 %. As we will see later, this would mean sending each month 40 references on an average, 14 of which would be relevant. In spite of the fact that NASA, ESRO and MEDLARS are serving their customers in this manner in my opinion such a service is not much appealing.

At the CID we were in a favorable situation: we had one year's experience of retrospective searches including feedback (that is important) before we started SDI. So we had time to think, and apply the results of the feedback study to SDI operation.

The following procedure is applied:

a) Each SDI search is handled at first like a retrospective search, but often restricted to documents published within the last few years. The customer is required to send feedback.

b) We do not start preparing the SDI profile until we have received the feedback of the retrospective search.

c) We assume that the customer will accept a certain amount of noise. A number of 20 irrelevant documents out of 100 sent could be acceptable, and the ratio could even increase for smaller shipments, because the customer's workload for checking the relevancy decreases.

   Figure 13 illustrates the "noise acceptance funktion". The 20 % noise in 100 references correspond to the starting point with 80 % relevancy. The 70 % noise for one reference per month means that the user is expected to accept 7 irrelevant documents out of 10 which he receives within 10 months.

d) By reformulating the queries we try to reduce the machine output to those references which we as well as the customer found to be relevant, without losing any relevant item, of course.

In terms of the graph that means that our operations should be represented by points below this curve. For most of the queries we were successful: point X indicates the values for an average SDI profile, the number of references sent being 19 to 20, the noise level being around 25 %, corresponding to a relevance ratio of 75 %. I do not know of any large documentation system in operation in which such a high relevancy is achieved without manual checking of the machine output and with practically no loss of relevant documents compared with retrospective searches.

How was it possible to increase the relevancy from 35 to 75 % without losing relevant documents? Here is the answer:

a) In a retrospective search we examine the machine response also to looser subqueries in order to make sure that no relevant documents are lost. Often we find no relevant documents at all; this reduces the overall relevancy for the query (covering all screened items) considerably.

b) After screening (that is visual checking) of the retrieved items, we often detect that some alternatives, even in relatively tight formulations, yield practically noise only. In the improved formulation we drop these alternative, or restrict them by further conditions, such as an additional AND-Operator.

c) In the screening process we often detect noise which could have been avoided by adding a group of terms with an AND NOT operator. The formulation can thus be improved, if we can make sure that no relevant items are lost.

d) We learn from the feedback which are the documents of real interest to the user; we can thus avoid retrieving such documents as are supplied after the retrospective search as "possible hits", just to be on the safe side.

e) During six years of indexing, the dictionary and the indexing rules were improved several times. For retrospective searches we have

to take into account, that indexing was not always done in the same way. In other words: The target is slightly diffuse, and we cannot focus our query exactly. This results in some noise. For SDI, on the contrary, we do not have to consider anything but the latest version of the dictionary and the rules, and we can rely on high consistency rates.

You see that there are some good reasons for the observed increase of the average relevancy from 35 to 75 %.

83 profiles were processed in March, 1968. There are a few without any response in March. For the broadest profile 260 references were retrieved. The average number of references per profile is 19.3. But we get a clearer picture by saying that for the first 90 % the average number of references is 9, whereas the average for the remaining 10 % is 114 references per profile. About 10 references per month is reasonable for an SDI service. The last 10 %, with hundreds of references, refer to documentation centers which redistribute the obtained references among their own customers.

The SDI relevancy figure of 75 % is our estimate. But which is the customer's opinion about the relevancy? I am in a position, fortunately, to answer this question: We received up to now 27 feedback letters on SDI, referring to more than 1,500 single references. These figures are brand new:

23 % of the references were pertinent and already known by the customer;

31 % of the references were pertinent and new to the customer;

20 % of the references were not just an answer to the query but useful;

26 % of the references were of no use (noise).

For an average query with 19 answer this would mean that 4 references are known by the user, 6 are not known and pertinent, 4 are not just pertinent but useful, and 5 are noise, as shown in figure 14.

You see, for SDI the customer almost precisely confirms our estimate of 75 % relevancy.

There remains one weak point in our system. That is the considerable time for screening, that is for manual checking of the references retrieved.

I remind you of the 3.5 hours required for the screening of an average of more than 100 abstracts per query. Screening is necessary because more than 60 references have to be discarded as noise, which corresponds to an average relevancy of 35 % for retrospective searches.

This effort is all the more unsatisfactory as it must be made by highly qualified staff, experienced not only in a scientific discipline but also in information processing, especially in indexing control, and retrieval query formulation. Furthermore, the customer will have to check the references anyhow. Now the job is done twice, at least for the pertinent and the possibly pertinent references.

Well, we found a solution to this problem, at least for queries with many answers in the output, and these are just those which make trouble:

1) We formulate a "medium query", not too tight and not too loose.

2) We select 10 to 20 abstracts and check their relevance.

3) We feed the abstract numbers into the computer, together with the indication "relevant" or "irrelevant". Up to this point it will be the documentalist's job. The rest will be left to the computer.

4) The computer extracts the index terms for the abstracts judged relevant and irrelevant.

5) The computer calculates a positive and a negative weight for the index terms, according to their percentage of occurence in the indexes of the relevant and the irrelevant abstracts, respectively. It then calculates the global weight as the difference of the positive and the negative weights.

6) The computer consults the indexes of the remaining items retrieved by the medium query and assigns to each term the weight calculated before.

ratio of 15.8 %. The 3 hits and 6 clearly non-relevant references, selected out of the 17 non- relevant references found in the sample, were used for the weighting and ranking procedure. The result is shown in the fourth line. The cut at 15 x 114 = 17 references yielded 11 hits (relevance ratio: 61 %).

The method is expected to increase the average relevancy from 35 to more than 80 %. If the customer is further interested in relevant references he may state in his comment to the literature search the relevance of the references he received. These statements will then again be fed into the computer to gain further improved output, and so on.

This sample screening and relevance feedback method is called PARIF: Program for Automatic Retrieval Improvement by Feedback. Our programmer started working on the program.

PARIF offers some possibilities for automatic indexing and retrieval. All non-trivial terms in the indexes of documents can be compared with all non-trivial terms of documents known as relevant or irrelevant, even without using a thesaurus but with weighting of terms and ranking of documents according to PARIF.

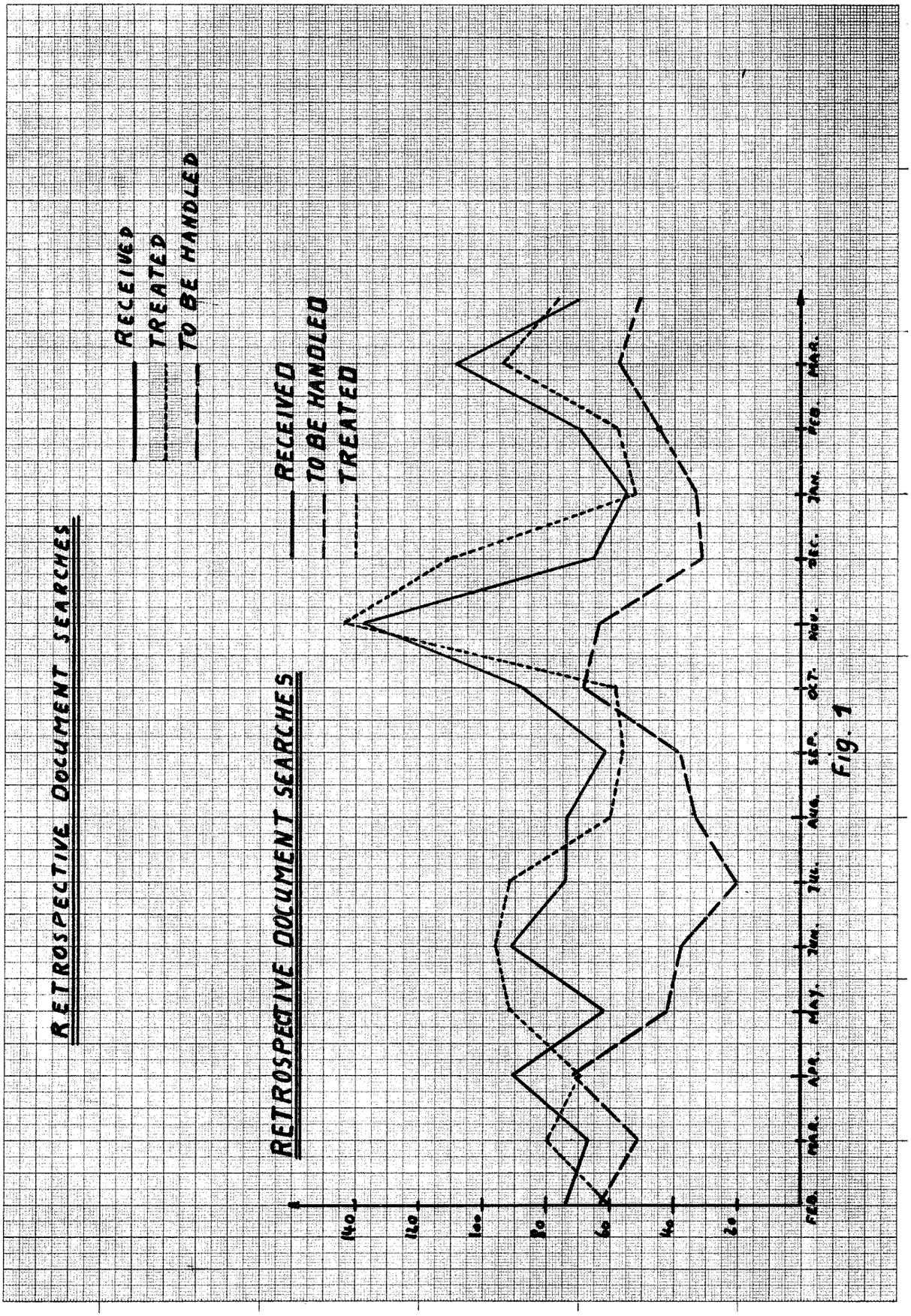

RETROSPECTIVE DOCUMENT SEARCHES
RECEIVED
TREATED
TO BE HANDLED
RETROSPECTIVE DOCUMENT SEARCHES
RECEIVED
TO BE HANDLED
TREATED
160
120
100
80
60
40
20
FEB. MAR. APR. MAY. JUN. JUL. AUG. SEP. OCT. NOV. DEC. JAN. FEB. MAR.

Fig. 1

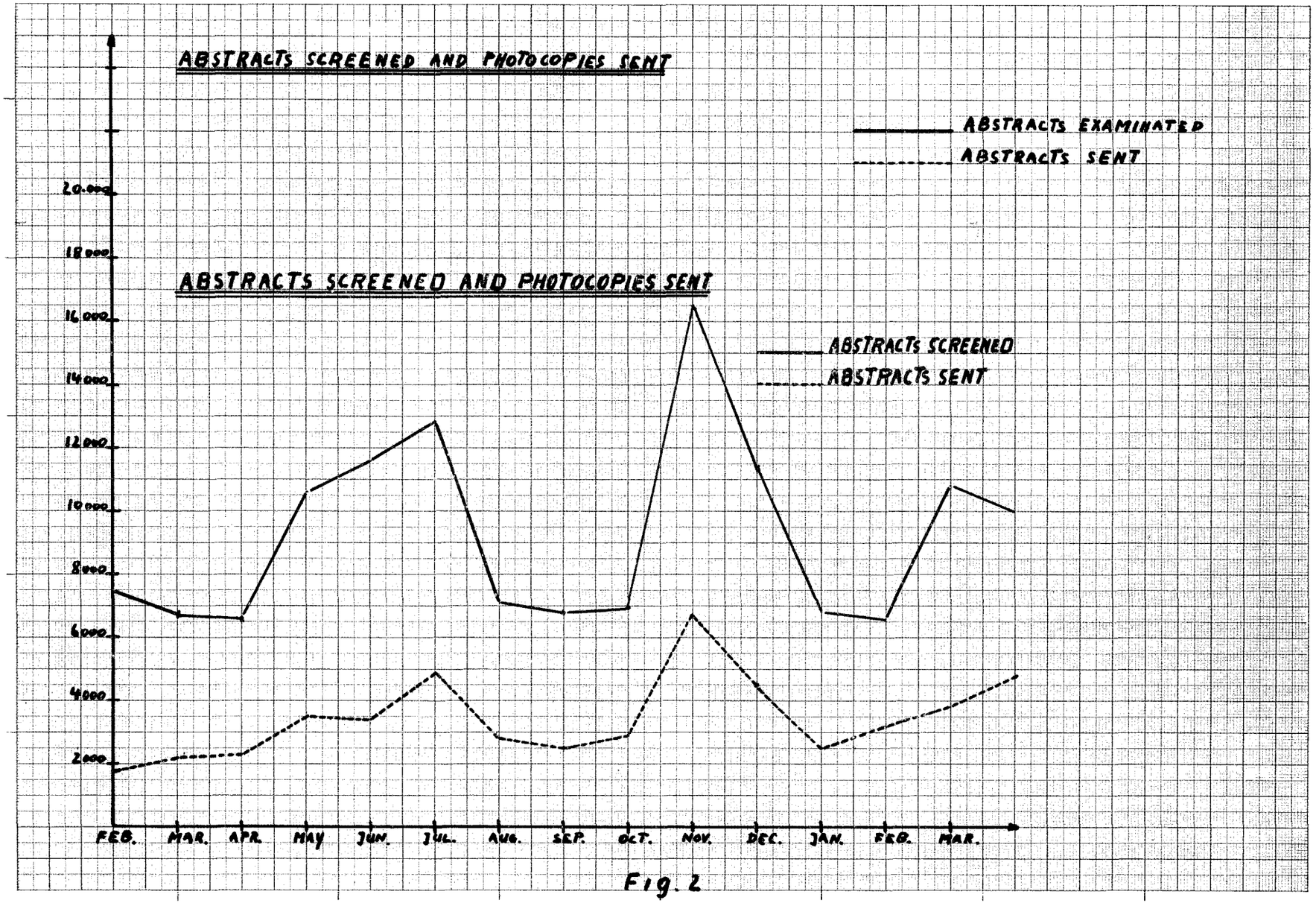

ABSTRACTS SCREENED AND PHOTOCOPIES SENT
ABSTRACTS EXAMINATED
ABSTRACTS SENT
20.000
18000
16,000
14000
12000
10000
8000
6000
4000
2000
FEB.
MAR.
APR.
MAY
JUN.
JUL.
AUG.
SEP.
OCT.
NOV.
DEC.
JAN.
FEB.
MAR.

Fig. 2

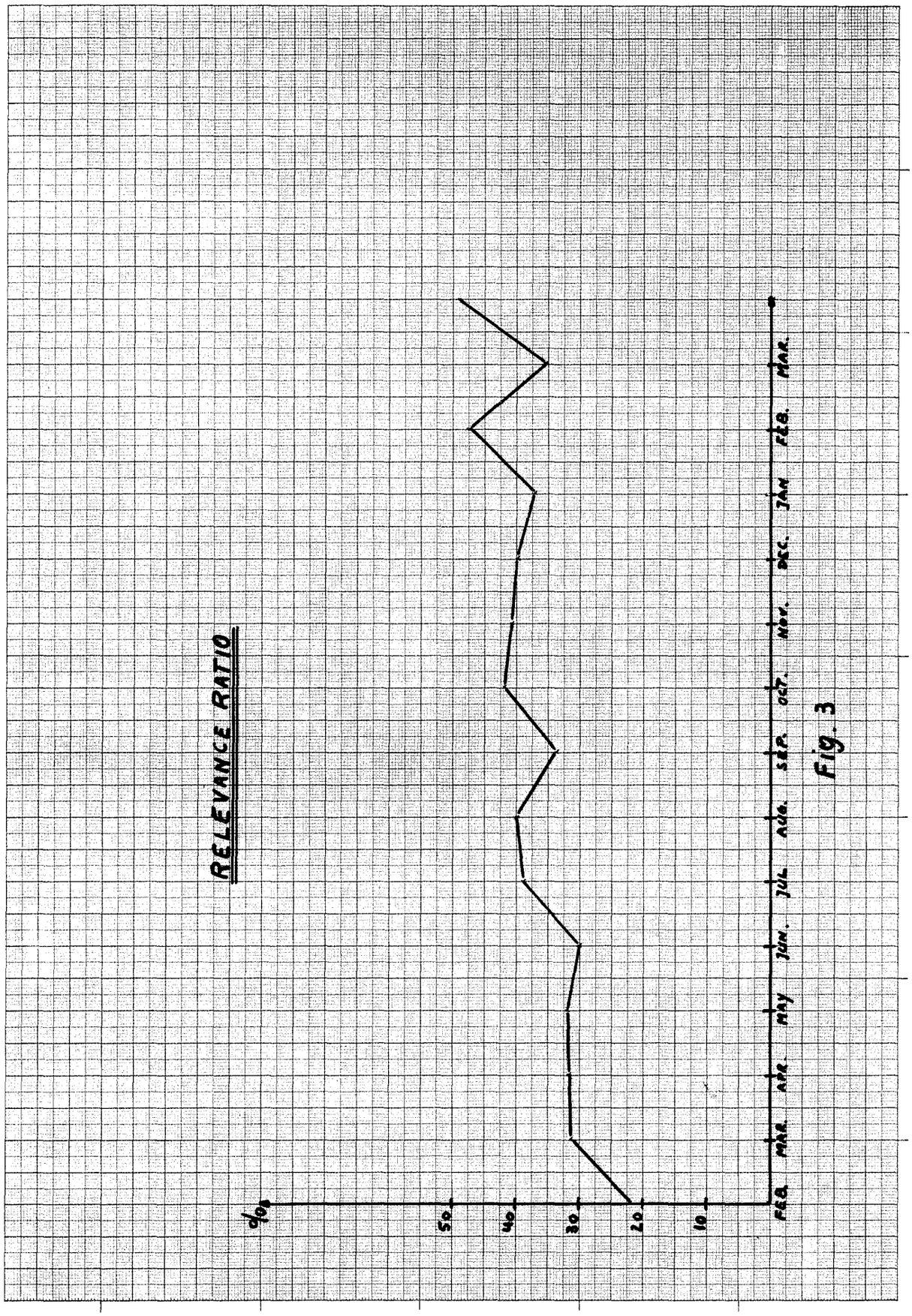

RELEVANCE RATIO
60
50
40
30
20
10
FEB. MAR. APR. MAY JUN. JUL. AUG. SEP. OCT. NOV. DEC. JAN. FEB. MAR.

Fig. 3

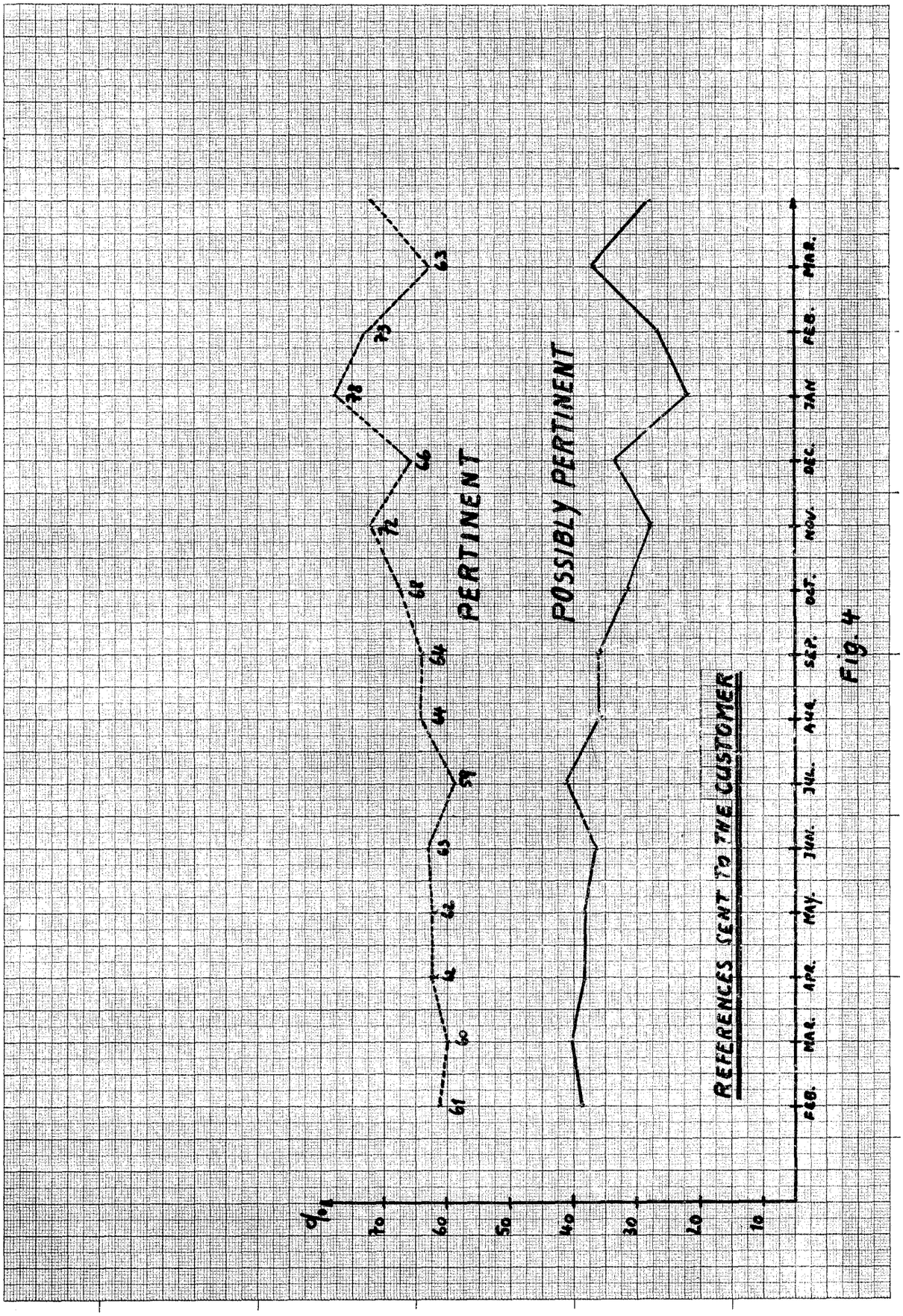

%
70
60
50
40
30
20
10
61
60
62
63
64
64
65
67
72
76
78
79
83
PERTINENT
POSSIBLY PERTINENT
REFERENCES SENT TO THE CUSTOMER
FEB.
MAR.
APR.
MAY.
JUN.
JUL.
AUG.
SEP.
OCT.
NOV.
DEC.
JAN
FEB.
MAR.

Fig. 4

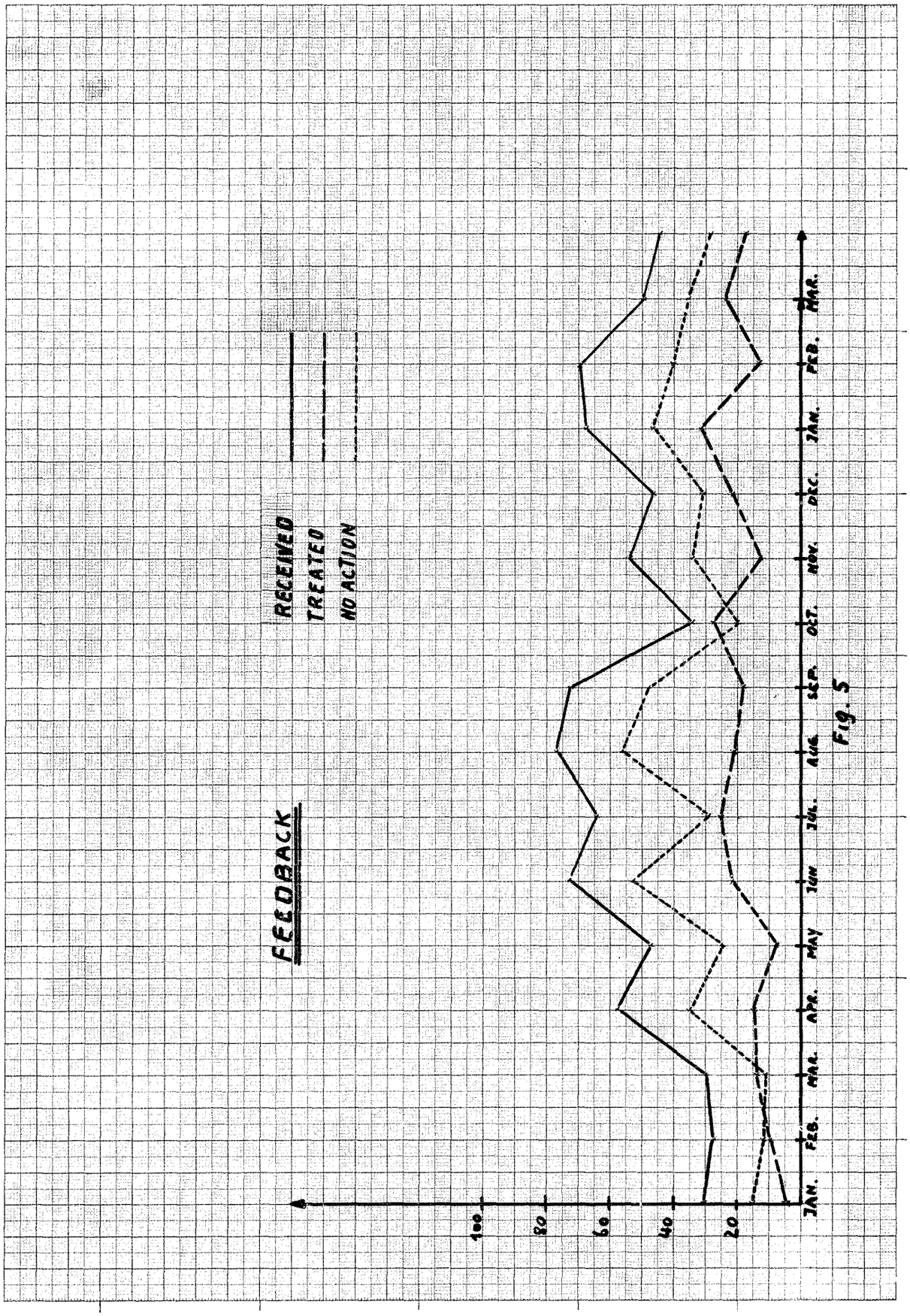

FEEDBACK
RECEIVED
TREATED
NO ACTION
100
80
60
40
20
JAN.
FEB.
MAR.
APR.
MAY
JUN.
JUL.
AUG.
SEP.
OCT.
NOV.
DEC.
JAN.
FEB.
MAR.

Fig. 5

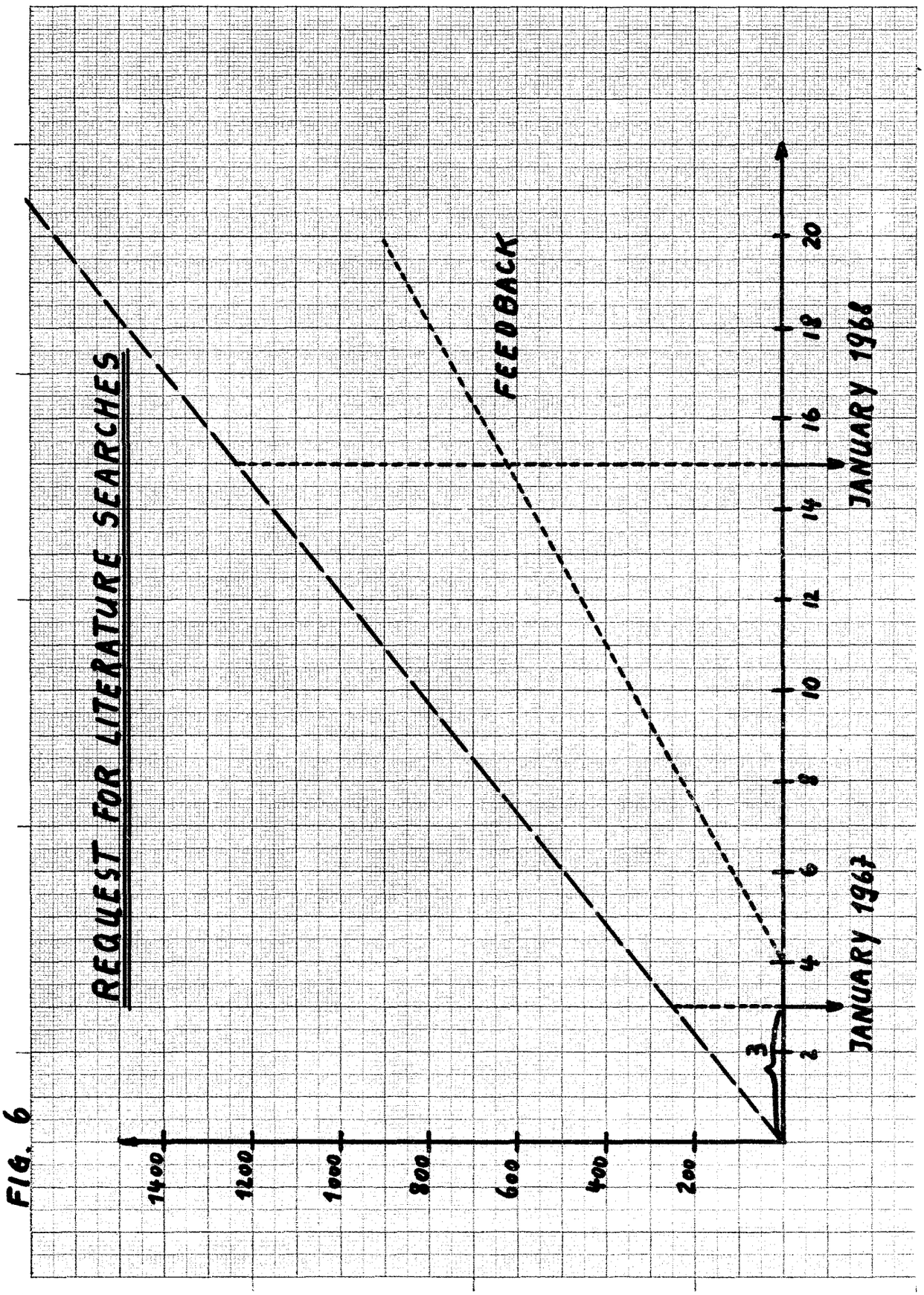

REQUEST FOR LITERATURE SEARCHES
FEEDBACK
1400
1200
1000
800
600
400
200
2
3
4
6
8
10
12
14
16
18
20
JANUARY 1963
JANUARY 1964

FIG. 6

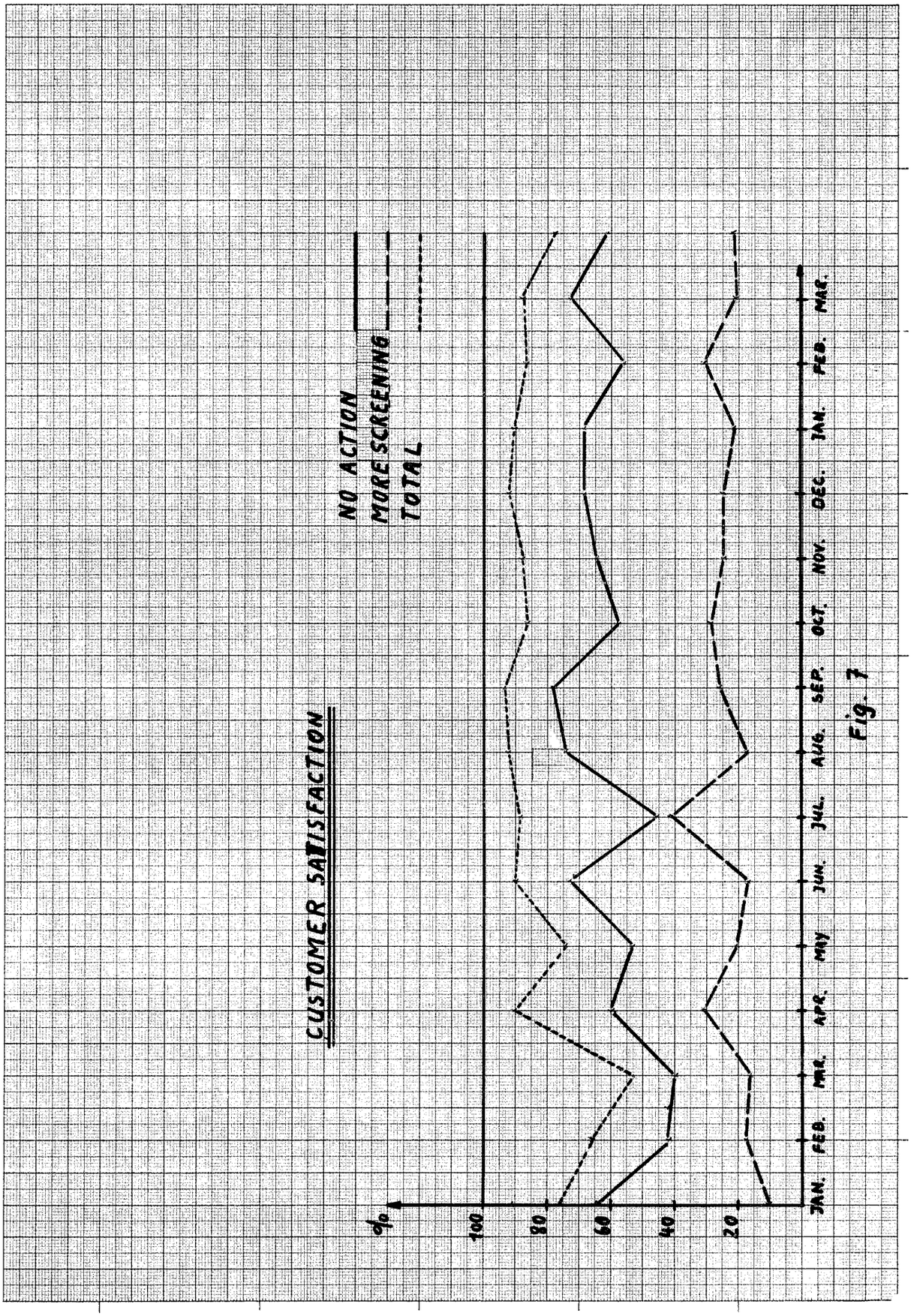
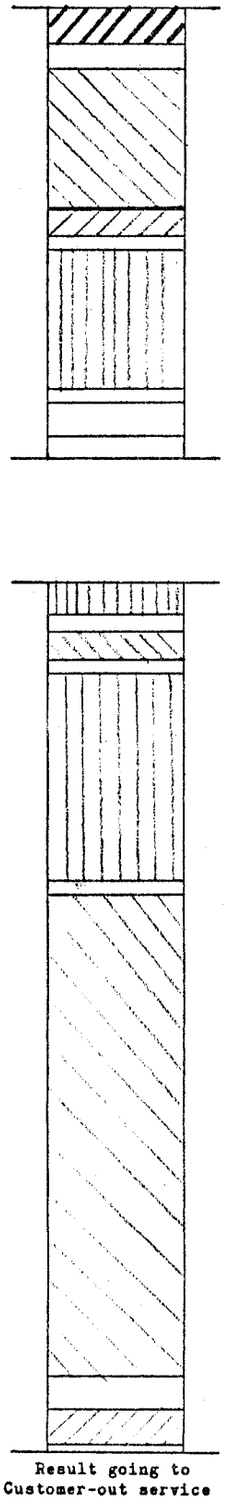

CUSTOMER SATISFACTION
NO ACTION
MORE SCREENING
TOTAL
% 100 80 60 40 20
JAN. FEB. MAR. APR. MAY JUN. JUL. AUG. SEP. OCT. NOV. DEC. JAN. FEB. MAR.

Fig. 7

Fig. 8

STEPS IN QUERY PROCESSING



15 min Check, precise and assign query to specialist!
10 min Register query; prepare Retrieval Sheet!

60 min Compare with older queries; formulate new query!

1 min Register query formulation!
12 min Check and improve query formulation!
6 min Register and pass for punching!

60 min Punch query formulations!

5 min Check completeness of punched card set!
15 min Checking of query formulation by the Computer ("Compiling")
10 min Correct errors!

50 min Retrieval run (at night)

batch of queries processed simultaneously

15 min Cut printout!
8 min Register, separate printout, pass for Screening Allocation!
12 min Indicate reference numbers in the printout ("Screening Allocation")!
3 min Register, pass to card file operators!

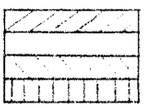
90 min Draw abstract cards!

6 min Register, pass for screening!

210 min Separate relevant from non-relevant references ("Screening")! fill in letter formsheet (draft)!

15 min Register, check completeness!

15 min Finally check complete search!
3 min Register, pass for photocopying relevant abstracts!



Head, Output Service
Secretary, Output Service
Subject Specialist
Operators (punching, machine processing, card drawing)

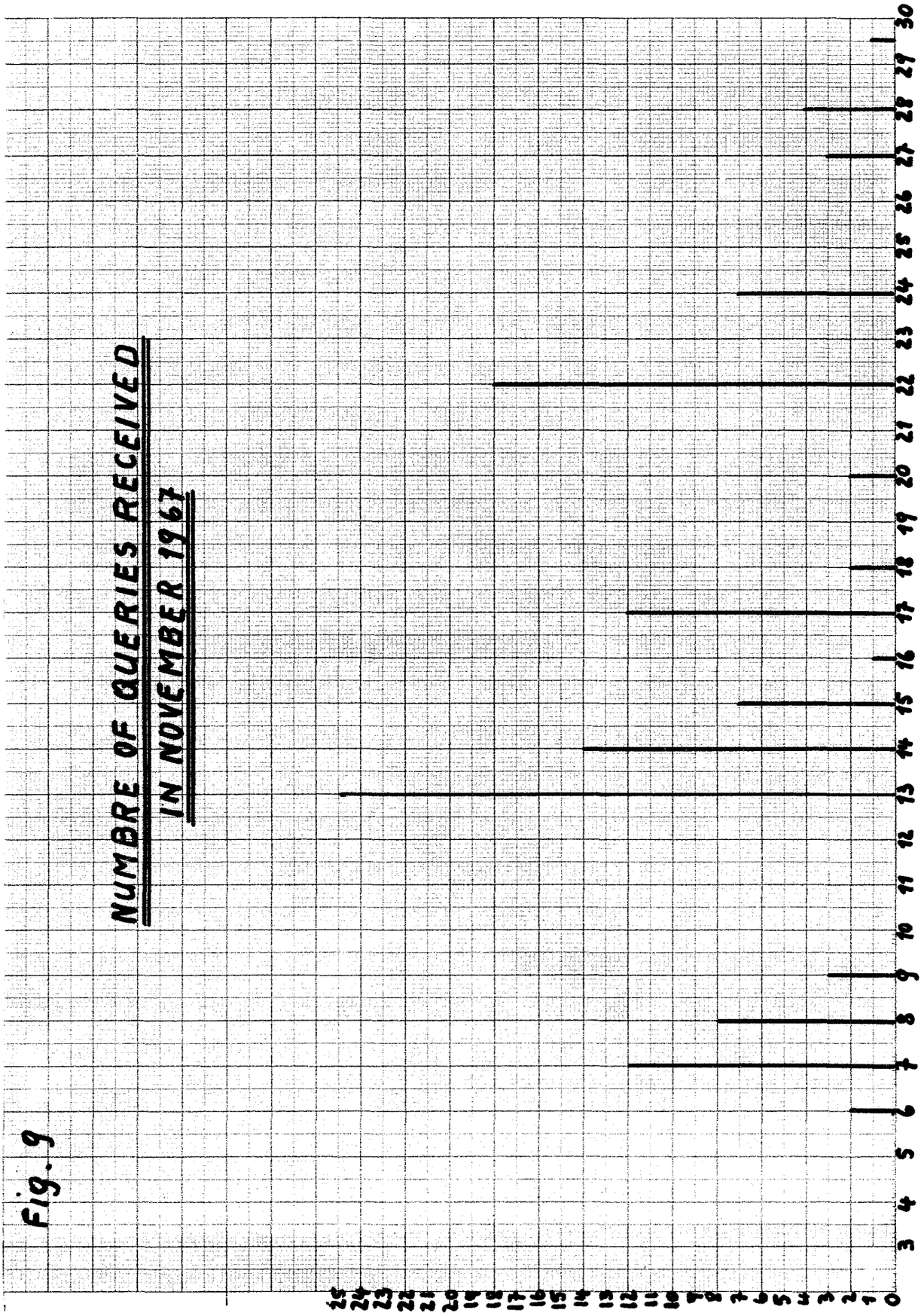

NUMBRE OF QUERIES RECEIVED
IN NOVEMBER 1967
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Fig. 9

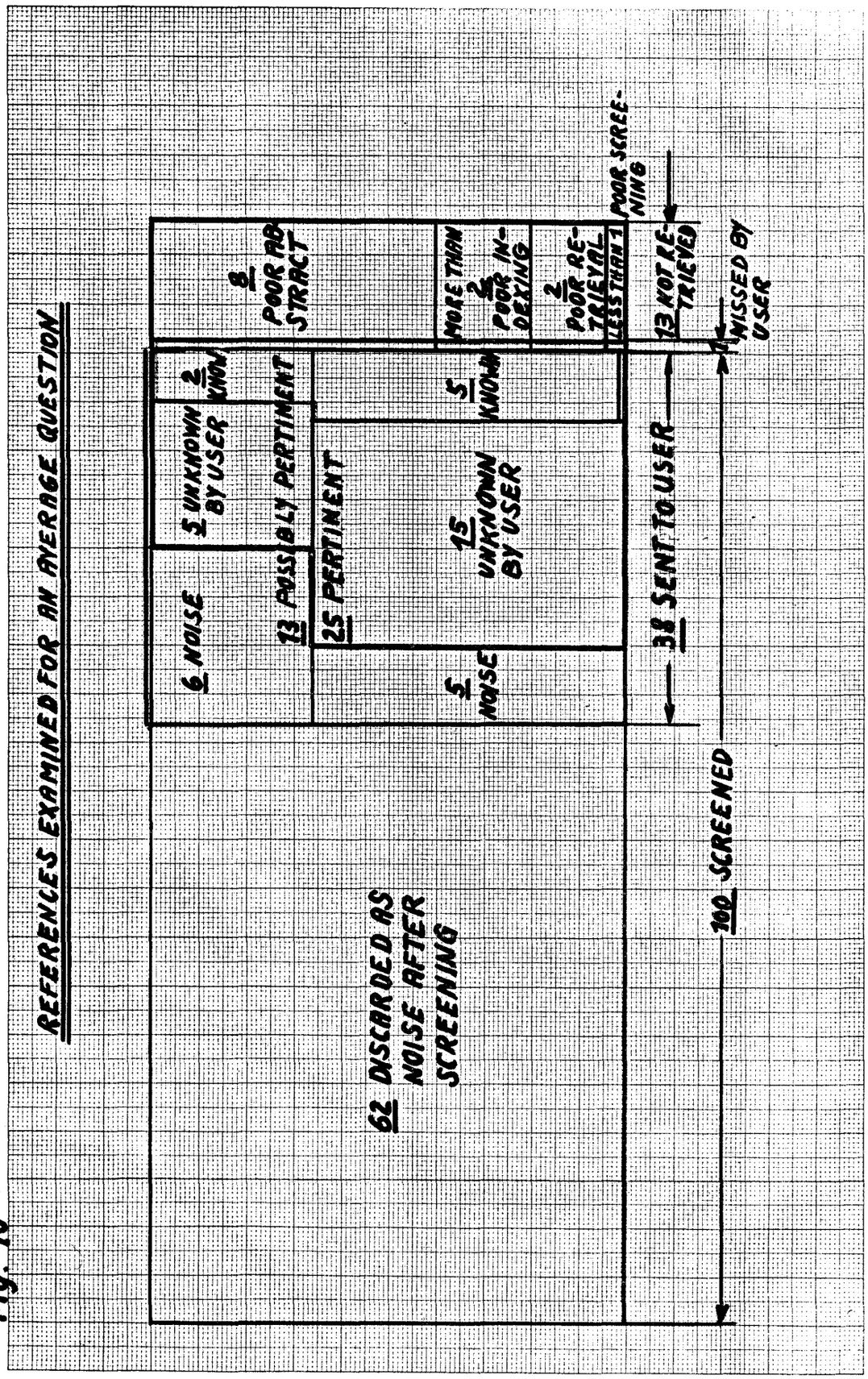

REFERENCES EXAMINED FOR AN AVERAGE QUESTION
100 SCREENED
62 DISCARDED AS NOISE AFTER SCREENING
38 SENT TO USER
MISSED BY USER
13 NOT RETRIEVED
POOR SCREENING
2 POOR RETRIEVAL LESS THAN 1
2 MORE THAN 2 POOR INDEXING
5 UNKNOWN
8 POOR ABSTRACT
5 NOISE
15 UNKNOWN BY USER
25 PERTINENT
13 POSSIBLY PERTINENT
2 UNKNOWN BY USER
6 NOISE

Fig. 10

Fig. 11

CUMULATIVE TOTAL OF CATEGORIES IN N S A
AND PERCENTAGE OF QUERIES

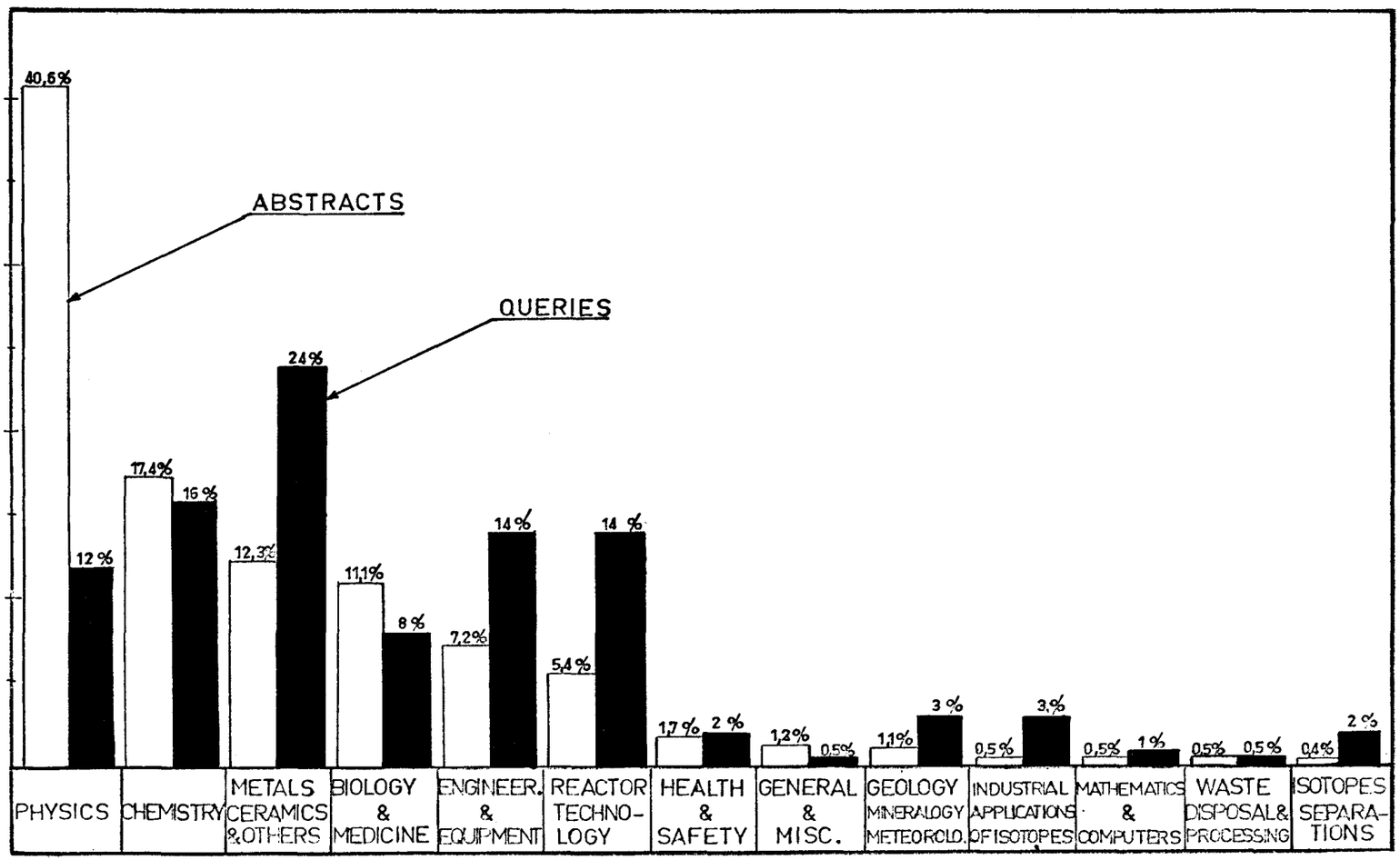


Fig. 12

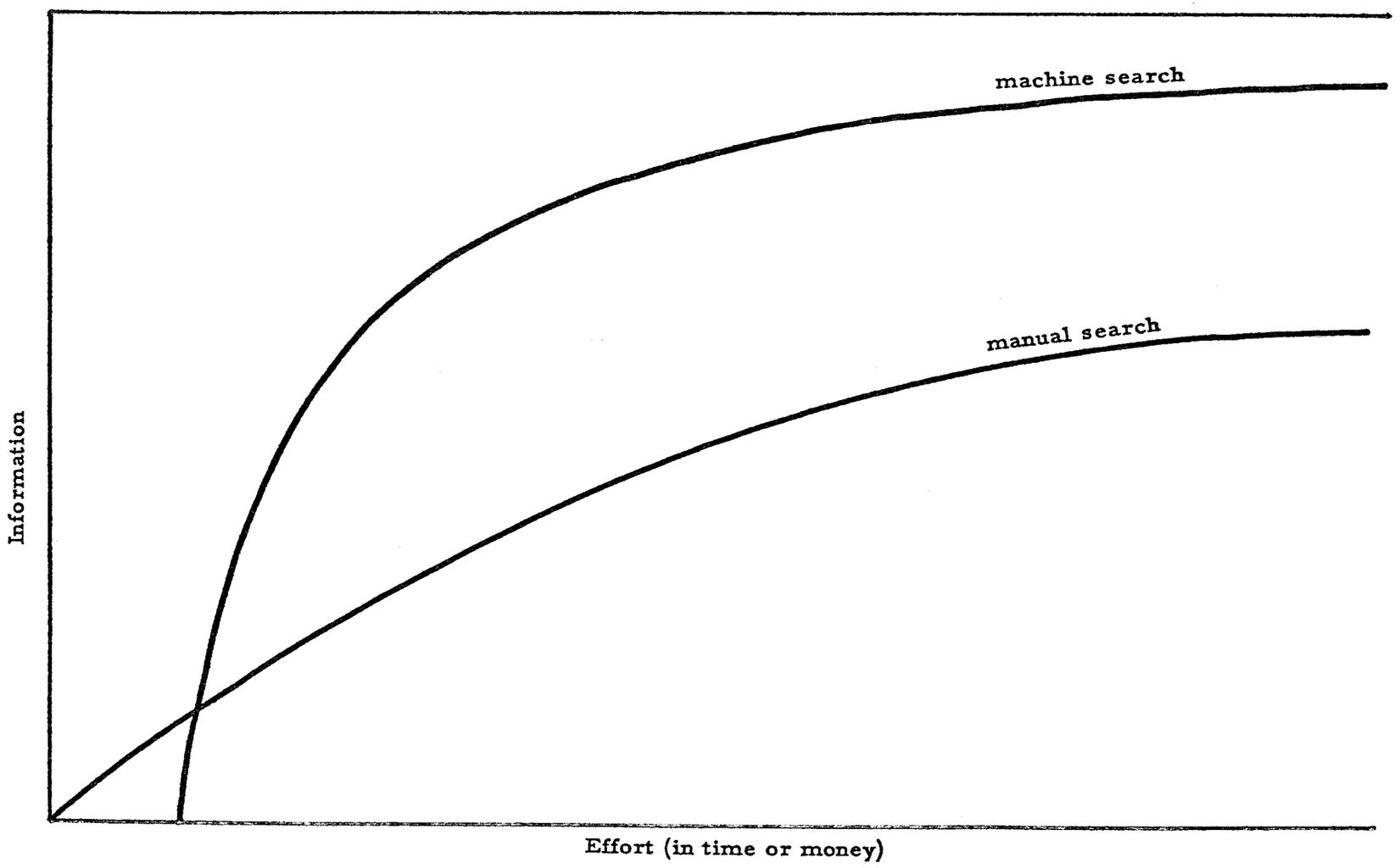
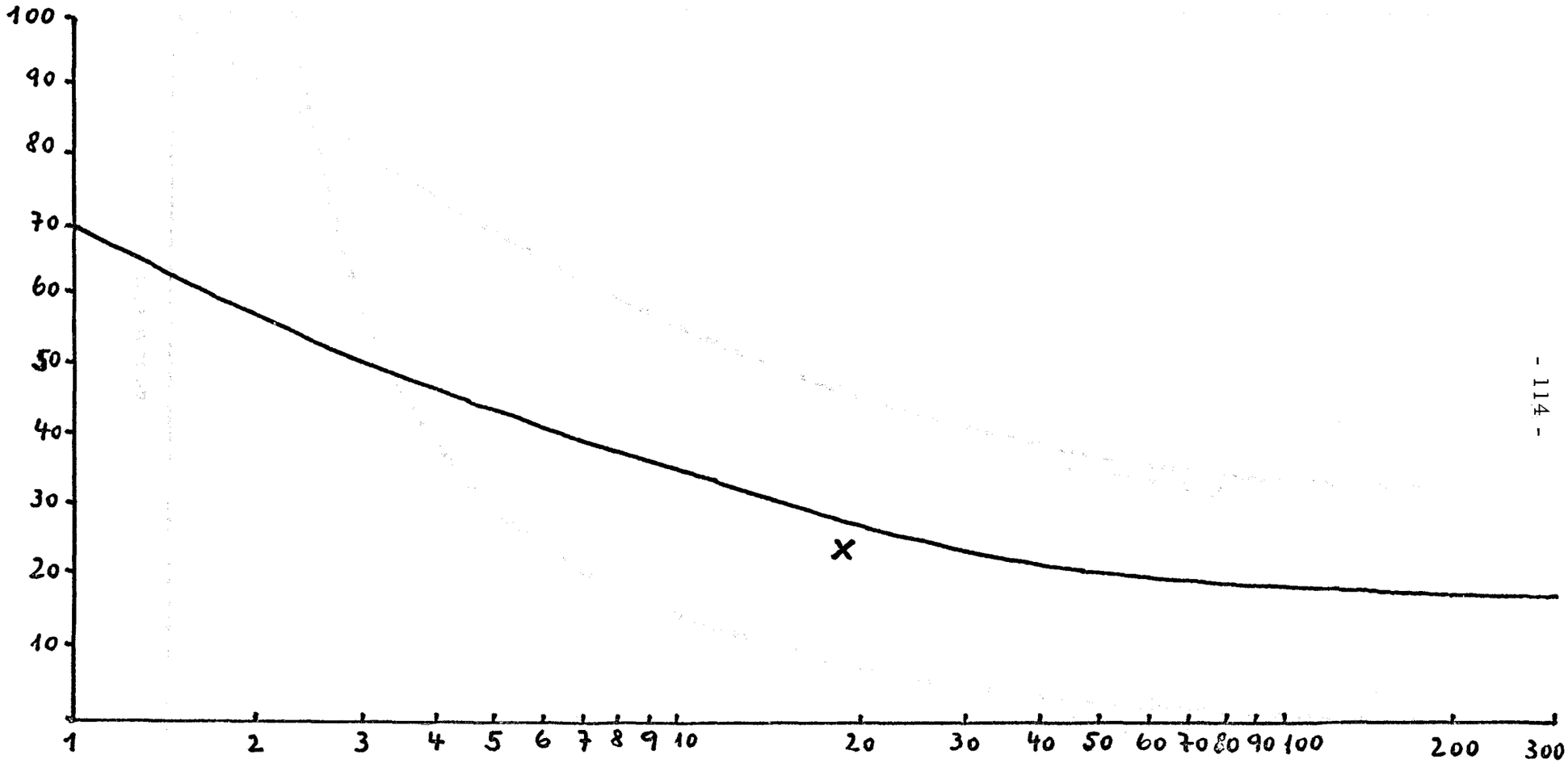

machine search
manual search
Information
Effort (in time or money)

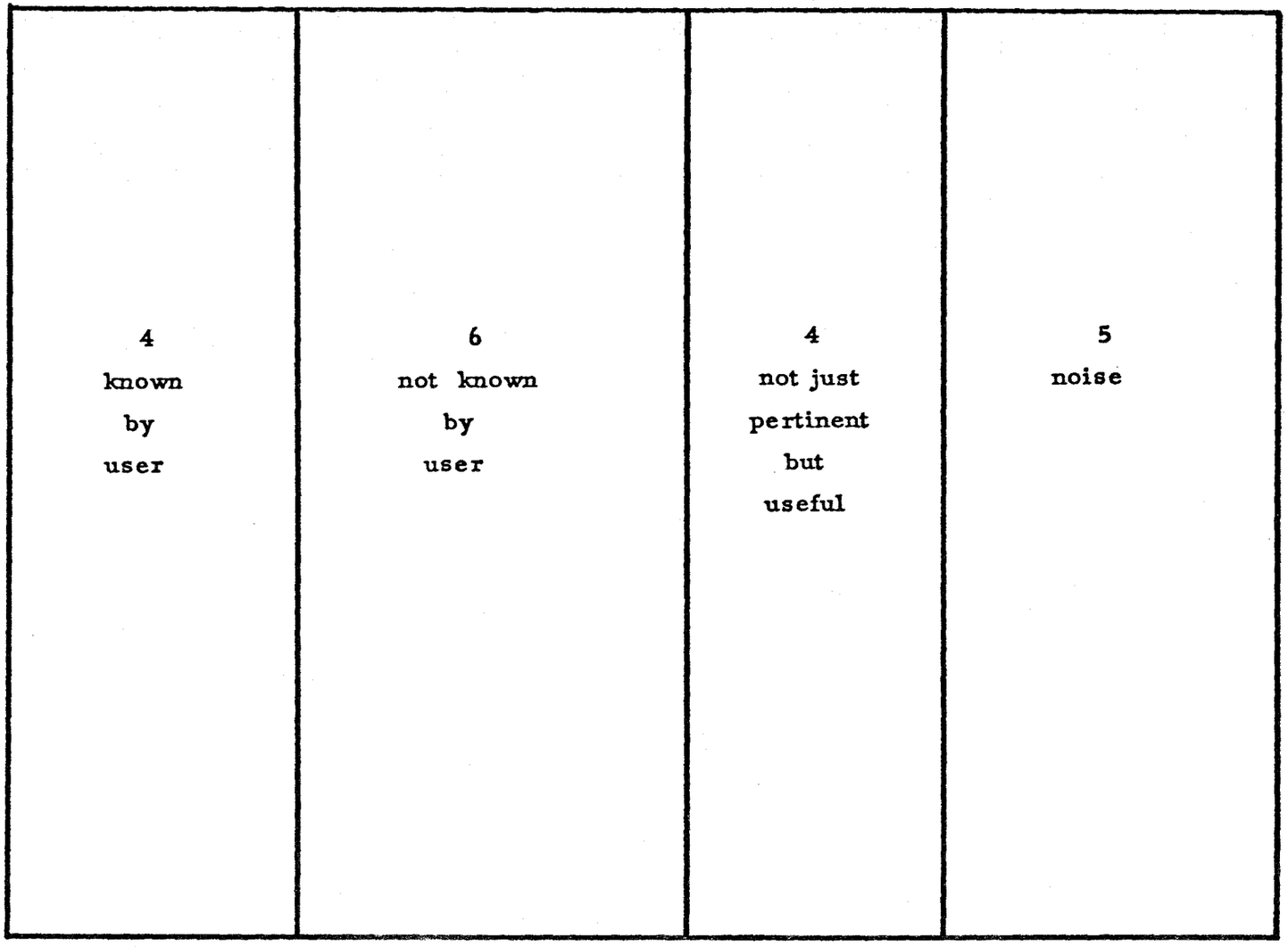
Fig. 13

NOISE ACCEPTANCE FUNCTION



NUMBER OF REFERENCES TO BE SENT TO THE CUSTOMER

Fig. 14



19 references sent per month for an average SDI query

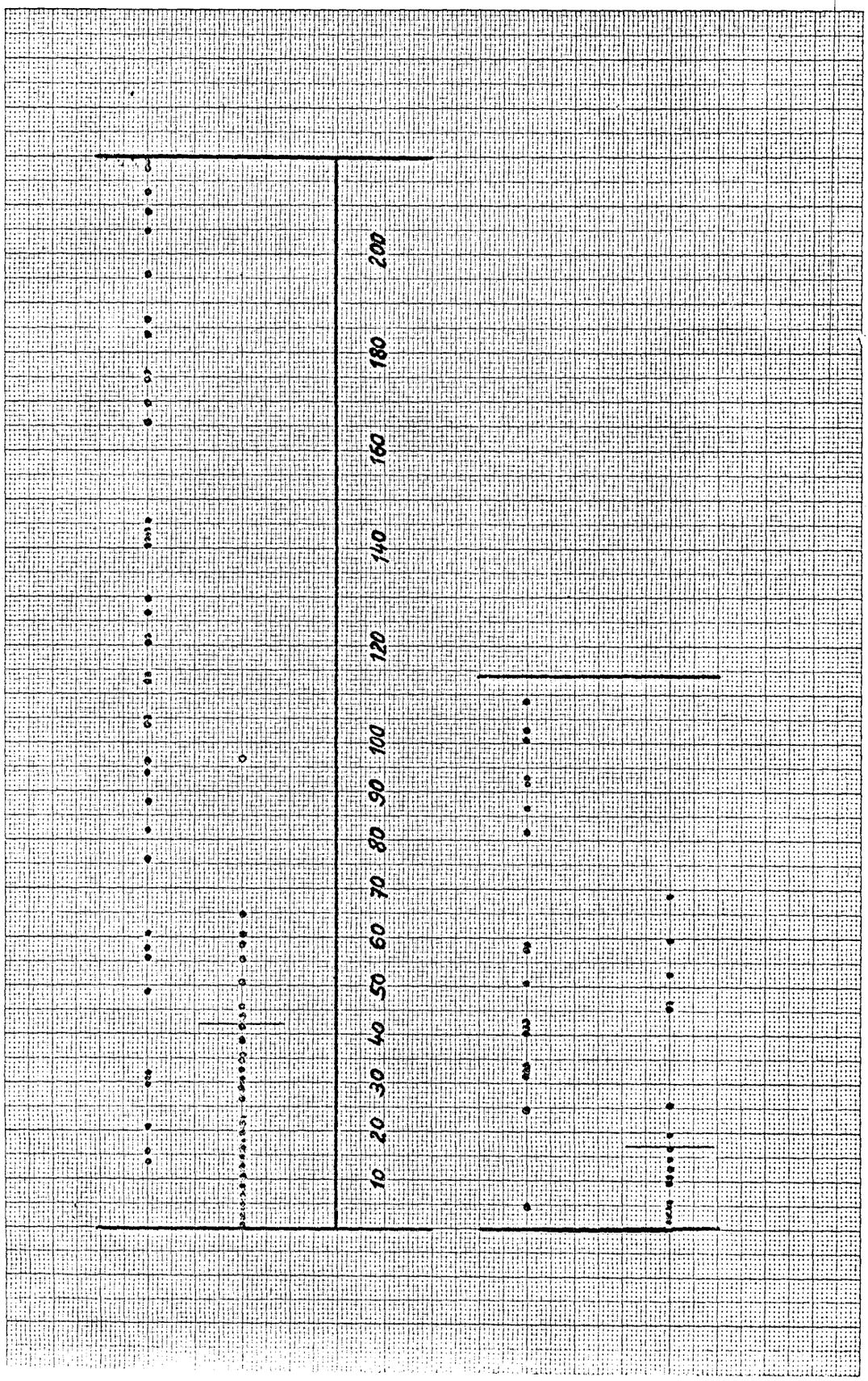


Fig. 15

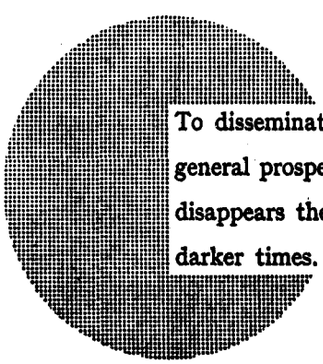
NOTICE TO THE READER

All Euratom reports are announced, as and when they are issued, in the monthly periodical **EURATOM INFORMATION,** edited by the Centre for Information and Documentation (CID). For subscription (1 year : US$ 15, £ 6.5) or free specimen copies please write to :

Handelsblatt GmbH
"Euratom Information"
Postfach 1102
D-4 Düsseldorf (Germany)

or

Centrale de vente des publications
des Communautés européennes
37, rue Glesener
Luxembourg



To disseminate knowledge is to disseminate prosperity — I mean general prosperity and not individual riches — and with prosperity disappears the greater part of the evil which is our heritage from darker times.

Alfred Nobel

SALES OFFICES

All Euratom reports are on sale at the offices listed below, at the prices given on the back of the front cover (when ordering, specify clearly the EUR number and the title of the report, which are shown on the front cover).

CENTRALE DE VENTE DES PUBLICATIONS DES COMMUNAUTES EUROPEENNES

37, rue Glesener, Luxembourg (Compte chèque postal N° 191-90)

BELGIQUE — BELGIË
MONITEUR BELGE
40-42, rue de Louvain - Bruxelles
BELGISCH STAATSBLAD
Leuvenseweg 40-42 - Brussel

DEUTSCHLAND
BUNDESANZEIGER
Postfach - Köln 1

FRANCE
SERVICE DE VENTE EN FRANCE
DES PUBLICATIONS DES
COMMUNAUTES EUROPEENNES
26, rue Desaix - Paris 15e

ITALIA
LIBRERIA DELLO STATO
Piazza G. Verdi, 10 - Roma

LUXEMBOURG
CENTRALE DE VENTE
DES PUBLICATIONS DES
COMMUNAUTES EUROPEENNES
37, rue Glesener - Luxembourg

NEDERLAND
STAATSDRUKKERIJ
Christoffel Plantijnstraat - Den Haag

UNITED KINGDOM
H. M. STATIONERY OFFICE
P. O. Box 569 - London S.E.1

EURATOM — C.I.D.
51-53, rue Belliard
Bruxelles (Belgique)