

EUR 4213 e

LIBRARY

EUROPEAN ATOMIC ENERGY COMMUNITY — EURATOM

**STATISTICAL ANALYSIS
OF
CONCEPT CO-ORDINATION
DOCUMENTATION SYSTEMS**

by

G.F. ROMERIO and A. CRICCHIO

1969



**Directorate-General for Dissemination of Knowledge
Center for Information and Documentation — CID**

AVERTISSEMENT

Le présent document a été élaboré sous les auspices de la Commission des Communautés Européennes.

Il est précisé que la Commission des Communautés Européennes, ses contractants, ou toute personne agissant en leur nom :

ne garantissent pas l'exactitude ou le caractère complet des informations contenues dans ce document, ni que l'utilisation d'une information, d'un équipement, d'une méthode ou d'un procédé quelconque décrits dans le présent document ne porte pas atteinte à des droits privatifs;

n'assument aucune responsabilité pour les dommages qui pourraient résulter de l'utilisation d'informations, d'équipements, de méthodes ou procédés décrits dans le présent document.

Ce rapport est vendu dans les bureaux de vente indiqués en 4^e page de couverture

au prix de FF 4,— FB 40,— DM 3,20 Lit. 500,— Fl. 3,—

Prière de mentionner, lors de toute commande, le numéro EUR et le titre qui figurent sur la couverture de chaque rapport.

Printed by Ceuterick — Louvain,
Brussels, January 1969.

EUR 4213 e

EUROPEAN ATOMIC ENERGY COMMUNITY — EURATOM

**STATISTICAL ANALYSIS
OF
CONCEPT CO-ORDINATION
DOCUMENTATION SYSTEMS**

by

G. F. ROMERIO and A. CRICCHIO

1969



**Directorate-General for Dissemination of Knowledge
Center for Information and Documentation — CID**

NOTARIAL PUBLIC STATE OF MISSISSIPPI

Know all men by these presents, that _____ of the County of _____ State of Mississippi, for and in consideration of the sum of _____ Dollars to _____ in hand paid by _____ the receipt of which is hereby acknowledged, have granted, sold and conveyed, and by these presents do grant, sell and convey unto the said _____ of the County of _____ State of Mississippi, all that certain _____

TO HAVE AND TO HOLD unto the said _____ heirs, assigns and assigns forever.

TO HAVE AND TO HOLD unto the said _____ heirs, assigns and assigns forever.

IN WITNESS WHEREOF, I have hereunto set my hand and seal of office this _____ day of _____ 19____.

SUMMARY

The basic principles for a statistical analysis of concept co-ordination documentation systems are outlined and developed starting from the representation matrix of the system.

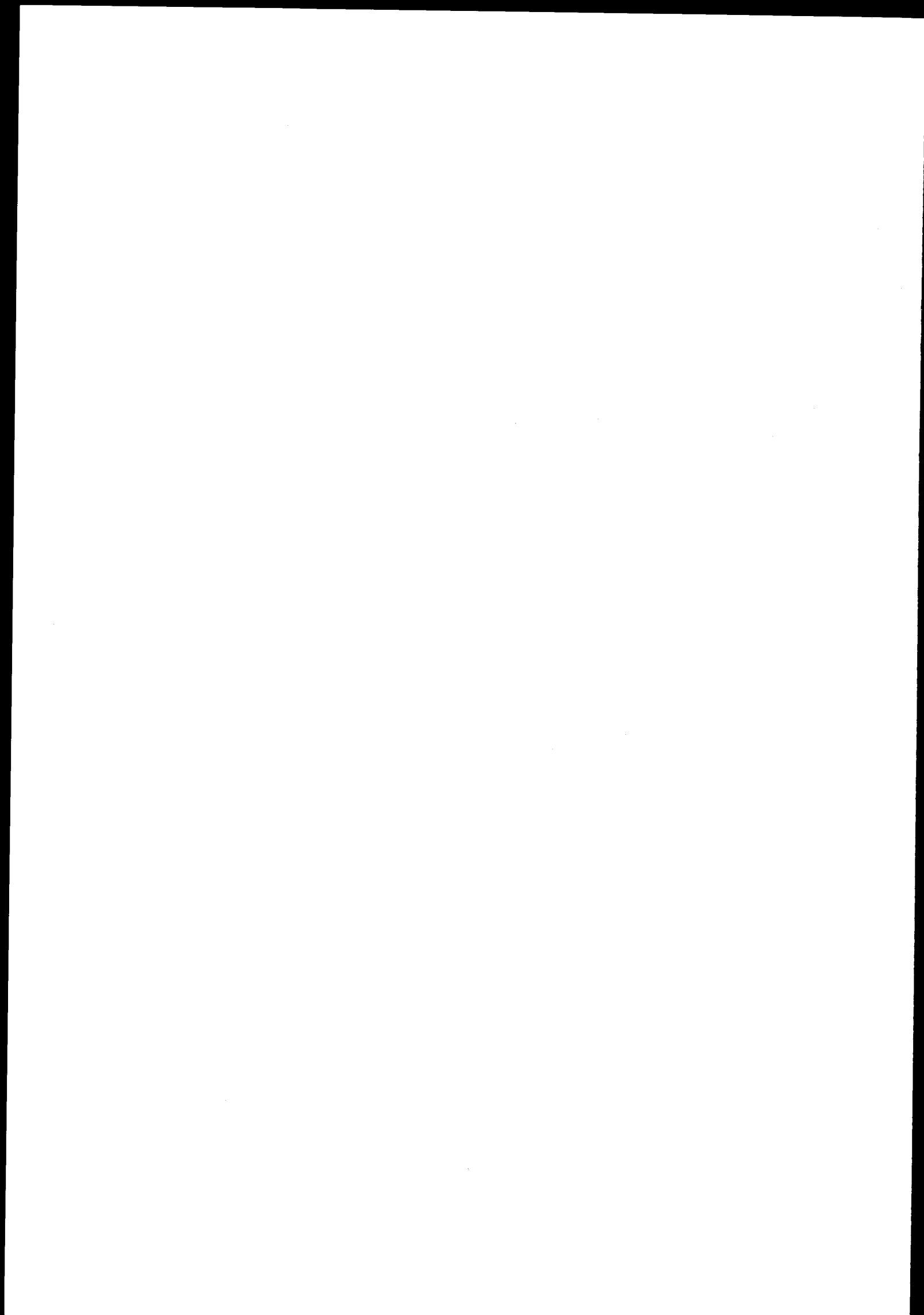
The concept of ideal log-normal system is introduced, where the distribution laws for the descriptors' assignment frequencies and for the documents' co-ordination levels are both log-normal with the same median.

The statistical analysis of EURATOM/CID Nuclear Documentation System is then performed and its characteristic parameters measured.

Finally, a study of the dynamics (i.e., the variation of the number of descriptors with the number of documents indexed) in a documentation system is carried out starting from a simple working hypothesis.

KEY WORDS

DOCUMENTATION
STATISTICS
MATRICES
EURATOM



INDEX

GLOSSARY	6
INTRODUCTION.	7
1 — REPRESENTATION MATRIX	7
2 — CONFIGURATIONS	9
3 — LOG-NORMAL SYSTEMS	11
4 — THE EURATOM/CID NUCLEAR DOC. SYSTEM.	14
5 — DYNAMICS OF DOCUMENTATION SYSTEMS	19
BIBLIOGRAPHY	25

GLOSSARY

a_1	number of descriptors introduced in the dictionary per any analyzed document
a_2	number of descriptors eliminated
A_f	$1/\sigma_f$
A_n	$1/\sigma_n$
B_f	$1/\sigma_f \log c_f$
B_n	$1/\sigma_n \log c_n$
c	median in a log-normal distribution
c_f	frequency median in log-normal systems
c_n	co-ordination level median in log-normal systems
C	number of complexions, i.e., of configurations of $\ R\ $
D_j	a document
f	attribution frequency, postings per descriptor
f_i	frequency of use of the descriptor W_i
f_0	average frequency
n	co-ordination level, postings per document
n_j	co-ordination level of the document D_j
n_0	average co-ordination level
N	total number of descriptors, dictionary size
N_0	dictionary size for $V = 0$
N_∞	asymptotic dictionary size
$p(i)$	posting probability for the i -th column of $\ R\ $
$p(j)$	posting probability for the j -th row of $\ R\ $
$p(f)$	probability density of f
$p(n)$	probability density of n
P	total number of postings in $\ R\ $
$r(f)$	rank of a descriptor with frequency f
$R(f)$	overtaking probability of f
$R(n)$	overtaking probability of n
$\ R\ $	representation matrix
s_0	selectivity of the dictionary, $s_0 = \log NV/P = \log V/f_0 = \log N/n_0$
t	a normalized variable $t = 1/\sigma_f \log f/c_f = 1/\sigma_n \log n/c_n$
V	total number of documents, data basis
V_c	a characteristic value of V , $V_c = 1/a_2$
W_i	a descriptor
ΔV	sample of V
ΔV_n	sample of V indexed at level n
λ	slenderness ratio, $\lambda = \log V/N$
σ	standard deviation in a gaussian distribution
σ_f	standard deviation for f in log-normal systems
σ_n	standard deviation for n in log-normal systems

STATISTICAL ANALYSIS OF CONCEPT CO-ORDINATION DOCUMENTATION SYSTEMS*

INTRODUCTION

The development and operation of large size documentation systems require that a solution should be found to the pressing problem of system optimization, paying attention to the well known system performance criteria as exhaustiveness, rapidity, precision, low cost and user appeal.

Some strategies studied to solve such a problem are based on information theory applied to documentation field. For every further development it is then necessary to know the structure of documentation systems both in their "static" behaviour, i.e. the situation at a certain moment, and in their "dynamics", i.e. the evolution during the growing up of the data basis. In particular the knowledge of the statistical distribution laws for descriptors' attribution frequencies and documents' indexing depths are of great importance. From information theory derives in fact that the information amount related to a descriptor is proportional to the logarithm of the reciprocal of its attribution probability.

It is interesting to observe also that the number of descriptors per document is a very significant pointer to the document importance (from a documentary and not scientific point of view). A document, in fact, to which only one or two descriptors have been attributed would be hardly ever retrieved, while on the other hand, a very deeply indexed document would cause often an increase of the noise in a documentary search.

In the following pages we try to justify the experimental statistical laws relating to the descriptors' attribution frequencies and documents' co-ordination levels, limiting our study to the very interesting case of a simple concept co-ordination system.

1 — REPRESENTATION MATRIX

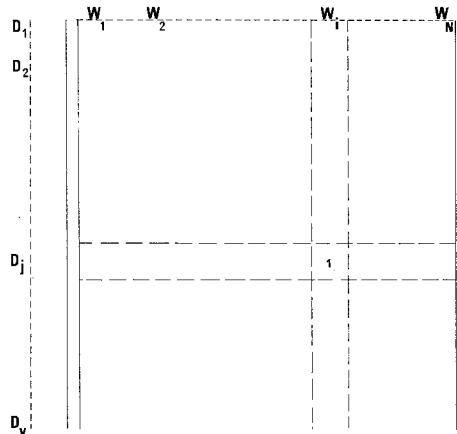
Basically, a documentation system consists of a collection of documents $[D_j]$ representing the data basis (messages) and a dictionary of descriptors, $[W_i]$, used to encode the documents. (See Note 1, page 24.)

If the system is of the "concept co-ordination" type, each document D_j is represented by a group of descriptors W_i which, through their association with D_j (co-ordination), characterize the contents of the document. If the descriptors are not interrelated (absence of links or roles) except by their simultaneous association with a given document, this is known as a simple co-ordination system. In this case it can be formally represented by means of a matrix $\|R\|$, called "representation matrix", in which each row corresponds to a document D_j and each column to a descriptor W_i . (See Note 2, page 24.)

In this matrix the posting of a descriptor W_i to the document D_j is represented by the figure 1 at the intersection of the i -th column and the j -th row. Hence the sum of the elements in the i -th column represents the frequency f_i of attribution of the descriptor W_i , i.e., the number of

* Manuscript received on September 24, 1968.

documents to which W_i is posted, whilst the sum of the elements in the j -th row represents the "co-ordination level" n_j of the document D_j , i.e., the number of descriptors posted to it.



Before embarking on the study of the structure of the matrix $\|R\|$ it is advisable to define some parameters that will occur in the coming developments.

If N is the number of descriptors in the dictionary and V the number of documents in the collection, the total number of postings P can be defined as follows:

$$P = \sum_1^N f_i = \sum_1^V n_j \quad (1)$$

From (1) we can derive the two following definitions, for the average attribution frequency f_0 and for the average co-ordination level n_0 (also called average indexing depth):

$$f_0 = \frac{P}{N} = \frac{\sum_1^N f_i}{N}; \quad n_0 = \frac{P}{V} = \frac{\sum_1^V n_j}{V}$$

whence:

$$P = f_0 N = n_0 V \quad (2)$$

Another parameter of interest is the logarithm of the ratio between the total number of documents and the total number of descriptors.

For this, taking (2) into account, we get:

$$\lambda = \log \frac{V}{N} = \log \frac{f_0}{n_0} \quad (3)$$

This parameter λ , which we shall call the matrix slenderness ratio, represents in a certain sense the redundancy of the system, i.e., it indicates the extent to which the documents deal with similar subjects. From the standpoint of system evolution, the matrix $\|R\|$ can be said to evolve according to a law by which λ increases with the number of documents in the collection ($d\lambda/dV > 0$). Let us now consider the density of posting, defined by the ratio

$$\frac{P}{NV} = \frac{n_0}{N} = \frac{f_0}{V}$$

which represents the probability that a certain posting ($W_i \rightarrow D_j$) is effected in respect of a certain element of the matrix. If the probability is low it can be said that the dictionary is very selective or of high definition, in that it contains a great many specific terms. The dictionary selectivity

in relation to the document collection can be expressed by the relation

$$S_0 = \log \frac{NV}{P} \quad (4)$$

or, using expression (2), by

$$S_0 = \log \frac{V}{f_0} = \log \frac{N}{n_0}.$$

2 — CONFIGURATIONS

From the purely analytical standpoint, the statistical study of the distribution of the co-ordination levels n and of the attribution frequencies f can be used to study the postings distribution in the representation matrix.

First we define the probabilities of posting

$$p(i) = \frac{f_i}{P} \quad p(j) = \frac{n_j}{P} \quad (5)$$

which represent the probabilities that the postings (the "1" of the matrix $\|R\|$) are located in the j -th row or the i -th column respectively. They make up a complete scheme, i.e., we have:

$$\begin{aligned} \sum_1^N p(i) &= \sum_1^N \frac{f_i}{P} = 1 \\ \sum_1^V p(j) &= \sum_1^V \frac{n_j}{P} = 1 \end{aligned} \quad (6)$$

Hence we can define the configuration of the matrix $\|R\|$ as the mode in which the whole P of the postings has been assigned. For instance, a configuration may have f_1 attributions in the first column, f_2 in the second and so on, subject to the condition

$$\sum_1^N f_i = f_1 + f_2 + \dots + f_N = P$$

We now calculate the total number C of possible configurations. This number (number of complexions) for the combinatorial analysis is given by the following:

$$C = \binom{P}{f_1} \cdot \binom{P-f_1}{f_2} \cdot \binom{P-f_1-f_2}{f_3} \cdot \dots \cdot \binom{P-f_1-f_2-\dots-f_{N-1}}{f_N}.$$

This expression can be simplified and expressed more compactly as:

$$C = \binom{P}{f_1 f_2 \dots f_N} = \frac{P!}{f_1! \cdot f_2! \dots f_N!} \quad (7)$$

i.e., by the well known multinomial coefficient.

The same reasoning can be used to calculate C from the attributions by row instead of by column. In this case we have

$$C = \binom{P}{n_1 n_2 \dots n_V} = \frac{P!}{n_1! \cdot n_2! \dots n_V!} \quad (8)$$

but (7) and (8) must be identically equal, i.e., we must have:

$$\frac{P!}{f_1! \cdot f_2! \dots f_N!} = \frac{P!}{n_1! \cdot n_2! \dots n_V!}$$

or

$$\prod_1^N (f_i!) = \prod_1^V (n_j!) \quad (9)$$

Considering the logarithms of (9), we have

$$\log \prod_1^N (f_i!) = \log \prod_1^V (n_j!)$$

i.e.

$$\sum_1^N \log f_i! = \sum_1^V \log n_j!.$$

Expressing the factorial by the Stirling formula

$$\log x! = (x + \frac{1}{2}) \log x - x + \frac{1}{2} \log 2\pi + \frac{r(x)}{12x} = x \log x - x + \frac{1}{2} \log 2\pi x + \left(\frac{r(x)}{12x}\right) \approx 0$$

we get

$$\sum_1^N f_i \log f_i - \sum_1^N f_i + \frac{1}{2} \sum_1^N \log 2\pi f_i = \sum_1^V n_j \log n_j - \sum_1^V n_j + \frac{1}{2} \sum_1^V \log 2\pi n_j.$$

It will be remembered that

$$\sum_1^N f_i = \sum_1^V n_j = P$$

we have

$$\sum_1^N f_i \log f_i + \frac{1}{2} \sum_1^N \log 2\pi f_i = \sum_1^V n_j \log n_j + \frac{1}{2} \sum_1^V \log 2\pi n_j \quad (10)$$

Taking (1) into account, equation (10) becomes

$$\frac{\sum_1^N f_i \log f_i}{\sum_1^N f_i} + \frac{\frac{1}{2} \sum_1^N \log 2\pi f_i}{\sum_1^N f_i} = \frac{\sum_1^V n_j \log n_j}{\sum_1^V n_j} + \frac{\frac{1}{2} \sum_1^V \log 2\pi n_j}{\sum_1^V n_j}.$$

It is quickly apparent that this expression contains the logarithmic means

$$\overline{\log f_i} = \frac{\sum_1^N f_i \log f_i}{\sum_1^N f_i}$$

$$\overline{\log n_j} = \frac{\sum_1^V n_j \log n_j}{\sum_1^V n_j}.$$

If the two terms

$$\frac{1}{2P} \sum_1^N \log 2\pi f_i ; \quad \frac{1}{2P} \sum_1^V \log 2\pi n_j$$

are negligible, we have

$$\overline{\log f_i} = \overline{\log n_j} = \overline{\log c} \quad (11)$$

Expression (10) becomes more nearly exact as the substitution

$$x! = x^x e^{-x}$$

gains in validity, i.e., the more x is large with respect to unity. Systems where (11) is exactly correct are called ideal systems.

3 — LOG-NORMAL SYSTEMS

The equations (11) in the foregoing paragraph show the logarithmic means of the descriptor attribution frequencies and document co-ordination levels weighted by the probabilities of posting $p(j)$ and $p(i)$. Let us now define the probabilities $p(n)$ and $p(f)$ as follows. Taking ΔN_f as the number of descriptors (matrix columns) with attribution frequency f , and ΔV_n as the number of documents (matrix rows) with level n , we write

$$\begin{aligned} p(f) &= \Delta N_f / N \\ p(n) &= \Delta V_n / V \end{aligned} \quad (13)$$

whence, through the logarithmic means, the following expressions are obtained

$$\begin{aligned} \log c_f &= \overline{(\log f)} = \sum_1^{f_{\max}} p(f) \cdot \log f \\ \log c_n &= \overline{(\log n)} = \sum_1^{n_{\max}} p(n) \cdot \log n \end{aligned} \quad (14)$$

Since the quantities n and f vary discretely and the smallest possible variation is unity, the magnitude $p(f)$ represents the probability that the frequency lies between $f - \frac{1}{2}$ and $f + \frac{1}{2}$. Therefore it is also the probability density since it is the probability referred to a unit interval.

However, for the density, we have

$$p(x) dx = p(\log x) d(\log x) = p(\log x) \frac{dx}{x} \quad (15)$$

i.e.,

$$p(x) = \frac{1}{x} p(\log x).$$

Now let us consider the first of expressions (14); in the light of (15) it becomes

$$\sum_1^{f_{\max}} p(f) \log f = \sum_1^{f_{\max}} \frac{1}{x} p(\log f) \log f = \overline{(\log f)}$$

which is approximately equal to:

$$\int_0^{f_{\max}} p(\log f) \log f d(\log f) = \log c_f$$

that is

$$\int_0^{f_{\max}} p(\log f) \log f \frac{df}{f} = \overline{\log f} = \log c_f$$

But we know that

$$\int_0^{f_{\max}} p(\log f) \log c_f d(\log f) = \log c_f$$

Subtracting the two preceding equations from one another, we obtain

$$\int_0^{f_{\max}} p(\log f) (\log f - \log c_f) d(\log f) = 0$$

that is,

$$\int_0^{f_{\max}} p\left(\log \frac{f}{c_f}\right) \cdot \log \frac{f}{c_f} \cdot d\left(\log \frac{f}{c_f}\right) = 0$$

Taking

$$\log \frac{f}{c_f} = u \quad ; \quad \log \frac{f_{\max}}{c_f} = u_{\max}$$

we get

$$\int_{-\infty}^{u_{\max}} p(u) \cdot u \, du = 0 \quad (16)$$

In (16) the integral obviously depends on the integration limits, whilst the second member is constantly zero. What should $p(u)$ be to make the integral constantly zero irrespective of u_{\max} ?

A sufficient condition can be found by resorting to an artifice.

Let us assume that the function $p(\log f/c_f)$ is regular. Since it must always be positive and run to zero at the end points of the definition interval

$$p\left(\log \frac{f_{\max}}{c_f}\right) = p\left(\log \frac{f_{\min}}{c_f}\right) \simeq 0$$

But then we have:

$$\int_0^{f_{\max}} d\left\{p\left(\log \frac{f}{c_f}\right)\right\} = 0$$

i.e., by putting $\log f/c_f = u$, we can write

$$\int_{-\infty}^{u_{\max}} d[p(u)] = 0 \quad (17)$$

Now let us consider the two integrals (16) and (17) and try to find a sufficient condition such that both are zero, irrespective of the limits of integration. To (16) we add (17) multiplied by an arbitrary constant σ^2

$$\int_{-\infty}^{u_{\max}} p(u) u \, du + \sigma^2 \int_{-\infty}^{u_{\max}} \frac{d[p(u)]}{du} \, du = 0$$

that is,

$$\int_{-\infty}^{u_{\max}} \left\{ p(u) u + \sigma^2 \frac{d[p(u)]}{du} \right\} du = 0. \quad (18)$$

For integral (18) to be equal to zero, irrespective of u_{\max} , i.e., of f_{\max} , it is sufficient that

$$p(u) u + \sigma^2 \frac{d[p(u)]}{du} = 0$$

By separating the variables we obtain

$$\frac{d[p(u)]}{p(u)} = -\frac{1}{\sigma^2} u du$$

i.e.,

$$d[\log p(u)] = -\frac{1}{2\sigma^2} d(u^2) = -d\left[\frac{u^2}{2\sigma^2}\right]$$

By resolving the foregoing differential equation and choosing a suitable normalization constant, we get

$$p(u) = \frac{1}{\sigma\sqrt{2\pi}} e^{-u^2/2\sigma^2} \quad (19)$$

In our developments we took $u = \log f/c_f$, but obviously the same reasoning can be used for the n values, so that (19) will still be valid with $u = \log n/c_n$.

Thus there exists a class of systems in which the condition (17) is fulfilled as to both f and n and in which the logarithmic means do not depend on the maximum values. For these systems the $p(f)$ and $p(n)$ distribution laws are of the type

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} e^{-(\log x/c)^2/2\sigma^2} \quad (20)$$

with

$$\begin{cases} x = f, n; & xp(x) = p(u) \\ \sigma = \sigma_f, \sigma_n \\ c = c_f, c_n \end{cases}$$

Expression (20) is the well-known Gibrat logarithmic-normal distribution. The systems in which it is valid are called log-normal.

With ideal log-normal systems, (20) is valid with $c_n = c_f = c$.

Then, as $p(x)$ is symmetrical with respect to $u = 0$, i.e., to $x = c$, c "characterizes" the mean, the median and the mode of the $p(x)$ distribution, i.e., the average value for which the integral probability is 1/2 and the most probable value.

Bearing in mind that

$$\int_0^x p(x) dx = \int_{-\infty}^u p(u) du$$

we find that the value $x = c$ (corresponding to $u = 0$) is the median of the distribution.

Concluding:

A. The ideal log-normal systems are characterized by the fact that the median of both distributions $p(n)$ and $p(f)$ is the same and coincides with the common value of the logarithmic means defined by (5) or by (14).

B. The average values of the magnitudes n and f , namely n_0 and f_0 , defined by (2), can be expressed through a property of the distribution (20) as a function of the constant c and of the dispersions σ_f and σ_n

$$\begin{aligned} n_0 &= c e^{\frac{1}{2}\sigma_n^2} \\ f_0 &= c e^{\frac{1}{2}\sigma_f^2} \end{aligned} \quad (21)$$

or :

$$\begin{aligned}\log n_0 &= \log c + \frac{1}{2}\sigma_n^2 \\ \log f_0 &= \log c + \frac{1}{2}\sigma_f^2\end{aligned}\quad (21')$$

From (21') we get

$$\log \frac{f_0}{n_0} = \frac{1}{2}(\sigma_f^2 - \sigma_n^2)$$

But through (3) we find

$$\log \frac{f_0}{n_0} = \log \frac{V}{N} = \lambda$$

so that

$$\lambda = \frac{1}{2}(\sigma_f^2 - \sigma_n^2) \quad (22)$$

C. In ideal log-normal systems the matrix slenderness ratio (system redundancy) is measured by the square deviation of the frequency dispersions and co-ordination levels. Then, since in a system σ_n^2 is constant, depending on the indexing "method", i.e., on n_0 ,

$$\sigma_f^2 = \sigma_n^2 + 2\lambda \quad (23)$$

This relation gives the frequency dispersion versus the matrix slenderness ratio (redundancy, "age" of system).

4 — THE EURATOM/CID NUCLEAR DOC. SYSTEM

Let us now, in the light of the foregoing considerations, look at the Euratom/CID Nuclear Documentation System. This is a simple co-ordinate indexing system (see para. 1) with a remarkably large collection of documents.

On January 1, 1968 the situation was as follows:

V	=	640,000 documents in the memory bank
N	=	12,066 dictionary descriptors
f_0	=	686 average frequency (attr./descriptors)
n_0	=	12.9 average coordination level (attr./doc.)
P	=	8,320,000 total number of postings

The statistical analysis of the co-ordination levels is effected by counting the number of documents $\Delta V_1, \Delta V_2, \dots, \Delta V_n$ that are encoded by 1, 2, ..., n descriptors respectively over a sufficiently representative sample.

If ΔV is the volume of the sample, the ratios $\Delta Vn/\Delta V$ approximate to the probabilities $p(n)$, if the sample is representative and ΔV is suitably large.

Table I shows the result of such a count effected on a sample of 10,000 documents.

The ratios $\Delta Vn/\Delta V$ are shown on the histogram in figure 1 in terms of n .

It is considerably harder to investigate the frequencies f on similar lines because the field of frequency variation ranges from 1 to about 100,000. But we can get round the practical impossibility of plotting an experimental histogram similar to the one in figure 1 by tackling the problem from another angle.

Let us suppose that the N descriptors are arranged in the opposite order to their frequency of attribution. We now assign a rank $r = 1$ to the most frequent descriptor, a rank $r = 2$ to the next most frequent, and so on, and try to graph the ranks r versus frequency.

TABLE I

Statistics of the co-ordination levels (indexing depths) performed on a sample $\Delta V = 10,000$ documents

n	ΔV_n	$p(n) = \frac{\Delta V_n}{\Delta V} \%$	n	ΔV_n	$p(n) = \frac{\Delta V_n}{\Delta V} \%$
1	70	0.7	19	300	3.0
2	100	1	20	190	1.9
3	120	1.2	21	210	2.1
4	220	2.2	22	180	1.8
5	280	2.8	23	120	1.2
6	440	4.4	24	110	1.1
7	670	6.7	25	80	0.8
8	560	5.6	26	110	1.1
9	760	7.6	27	90	0.9
10	810	8.1	28	60	0.6
11	870	8.7	29	80	0.8
12	620	6.2	30	80	0.8
13	530	5.3	31	80	0.8
14	540	5.4	32	30	0.3
15	470	4.7	33	20	0.2
16	370	3.7	34	40	0.4
17	490	4.9	35	10	0.1
18	300	3.0			

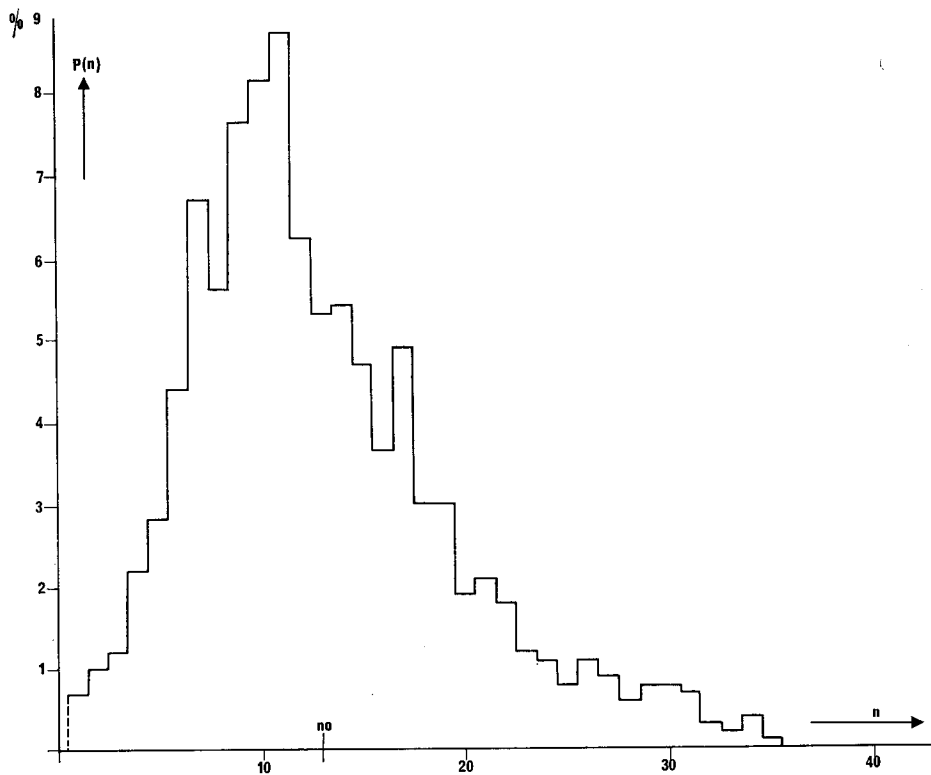


Fig. 1

We know that Zipf found, as to natural languages, that the rank of each word present in the texts examined is inversely proportional to that word's frequency of use.

This is expressed by Zipf's law

$$r(f) = K \frac{1}{f}. \quad (24)$$

If this is true, the graph of r versus f on log-log paper should be a straight line at 45° . But we also know that Zipf's law does not apply to documentation systems where the dictionary has no such redundancy as is found in spoken languages. Figure 2 shows the graph of the ranks (ordinates) versus the respective frequencies (abscissae) as it appeared from the data stored in the CID computer memory.

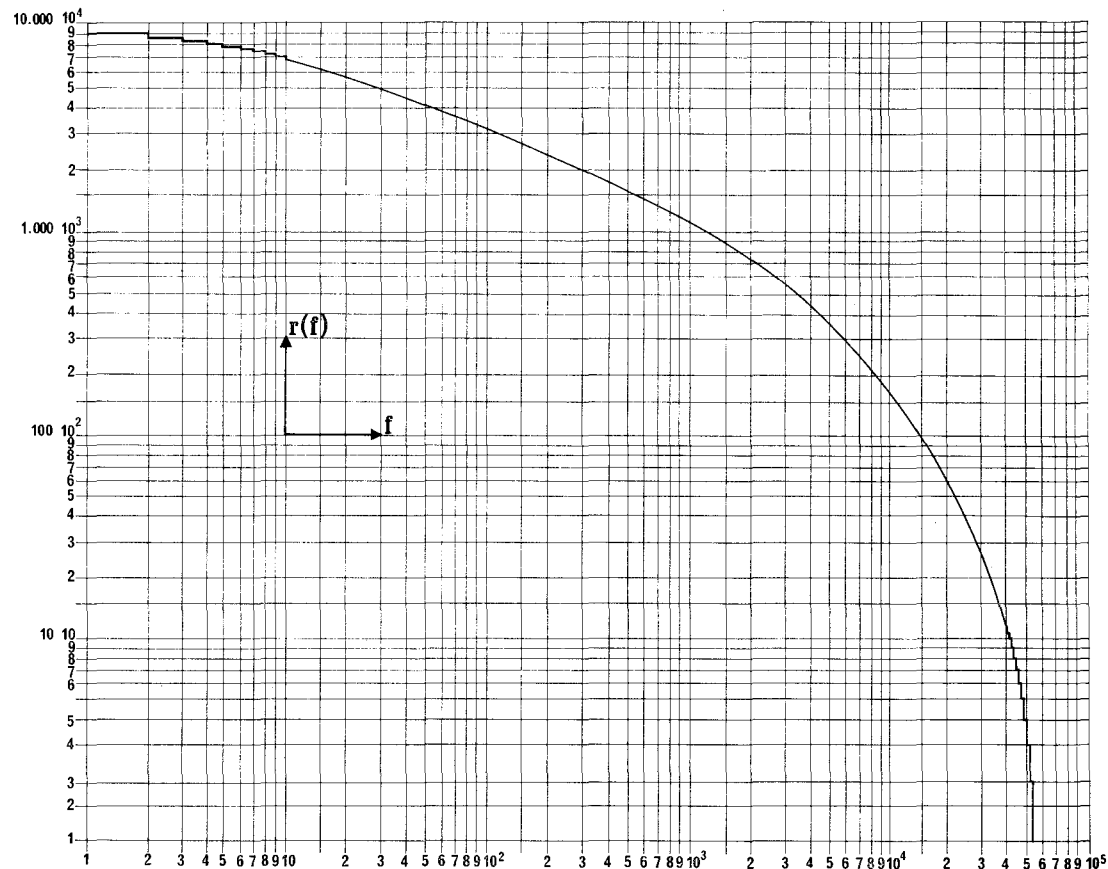


Fig. 2

As may be seen, the curve plotted on log-log graph paper is a long way from the law expressed by (24). Yet, as traced here, i.e., with the f and r axes inverted, it is of major importance to the statistical analysis that we had set ourselves to do.

For it need only be observed that, with a given frequency f , the corresponding rank $r(f)$ as read from the graph gives the number of descriptors whose frequency is greater than or equal to f (in discontinuity points the rank is the lower one). Hence the ratio $r(f)/N$ expresses the probability that, taking a descriptor at random, its attribution frequency will be greater than (or equal to) f .

Lastly, if we look once more at the histogram in figure 1, we can obtain the two following expressions of the (integral) overtaking probabilities:

$$R(n) = 1 - \sum_1^{\Delta V} \Delta V_n / \Delta V$$

$$R(f) = r(f)/N \quad (25)$$

Tables II and III give the values $R(n)$ and $R(f)$ calculated from (25).

TABLE II
Overtaking probabilities $R(n)$, calculated with the data of table II

n	$\Sigma \Delta V_n$	$\Delta V - \Sigma \Delta V_n$	$R(n) \%$
1	70	9 930	99.3
2	170	9 830	98.3
3	290	9 710	97.1
4	510	9 490	94.9
5	790	9 210	92.1
6	1 230	8 770	87.7
7	1 900	8 100	81.0
8	2 460	7 540	75.4
9	3 220	6 780	67.8
10	4 030	5 970	59.7
12	5 520	4 480	44.8
14	6 590	3 410	34.1
16	7 430	2 570	25.7
18	8 220	1 780	17.8
20	8 710	1 290	12.9
25	9 410	590	5.9
30	9 830	170	1.7
35	10 000	0	

We now propose to ascertain whether the two probability densities $p(n)$ and $p(f)$ corresponding to the overtaking probabilities $R(n)$ and $R(f)$ are log-normal.

If so, we can write
with $x = n, f$

$$p(x) = \frac{1}{\sigma_x \sqrt{2\pi}} \cdot \frac{1}{x} e^{-(\log x/c)^2 / 2\sigma^2}$$

Putting

$$t = \frac{1}{\sigma_x} \log \frac{x}{c}$$

and since

$$p(x) dx = p(t) dt \quad (26)$$

we obtain

$$\begin{cases} p(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \\ t = \frac{1}{\sigma_x} \log x - \frac{1}{\sigma_x} \lg c = A \lg x - B. \end{cases} \quad (27)$$

TABLE III
 Calculations of $R(f)$ from data of the frequency ranks (fig. 2)
 $N = 12,066$ descriptors

f	$r(f)$	$R(f) \%$	f	$r(f)$	$R(f) \%$
0	9 711	80.5	200	2 200	18.2
1	9 132	75.6	300	1 820	15.1
2	8 683	72.0	400	1 610	13.3
3	8 346	69.0	500	1 460	12.1
4	7 997	66.2	600	1 350	11.2
5	7 711	63.8	700	1 260	10.4
6	7 436	61.5	800	1 200	9.9
7	7 177	59.4	900	1 150	9.5
8	6 921	57.3	1 000	1 070	8.9
9	6 611	54.8	1 400	900	7.5
10	6 425	53.2	2 000	730	6.1
14	6 000	49.7	3 000	550	4.6
20	5 500	45.6	4 000	410	3.4
30	4 800	39.8	5 000	325	2.7
40	4 350	36.0	6 000	270	2.2
50	4 000	33.1	7 000	230	1.9
60	3 750	31.0	8 000	195	1.6
70	3 500	29.0	9 000	170	1.4
80	3 420	28.3	10 000	150	1.2
90	3 200	26.5	14 000	90	0.9
100	3 050	25.3	20 000	50	0.4
140	2 600	21.6	30 000	25	0.2
			40 000	9	0.09

The expression of the integral probabilities $R(n)$ and $R(t)$ will be

$$R(x) = 1 - \int_0^x p(x) dx$$

$$R(t) = 1 - \int_0^t p(t) dt = 1 - \text{erf}(t)$$

where $\text{erf}(t)$ is the Gauss error function.

Through (26) and (27) we finally obtain, if $p(x)$ is log-normal,

$$\left. \begin{aligned} R(x) &= 1 - \text{erf}(t) \\ t &= A \log x - B \end{aligned} \right\} \quad (28)$$

In our case, the checking of (28) against the values in tables II and III was done on log-normal paper. The resultant graph in figure 3 enables us to conclude that the Euratom/CID Nuclear Doc. System is, nearly enough, an ideal log-normal system. The two functions $R(n)$ and $R(f)$ can be approximated by means of two straight lines which fulfil condition (28) and intersect at a point where $R(n) = R(f) = 1/2$.

From the graph in figure 3 we obtain the following values:

$$\text{As to frequencies, } A_f = 0.35 \quad B_f = 0.85$$

$$\text{As to levels, } A_n = 1.96 \quad B_n = 4.70$$

Hence:

$$\log c = \frac{B_f}{A_f} = \frac{B_n}{A_n} = 2.42 \quad ; \quad c = 11.35$$

$$\sigma_f = 2.86$$

$$\sigma_n = 0.51$$

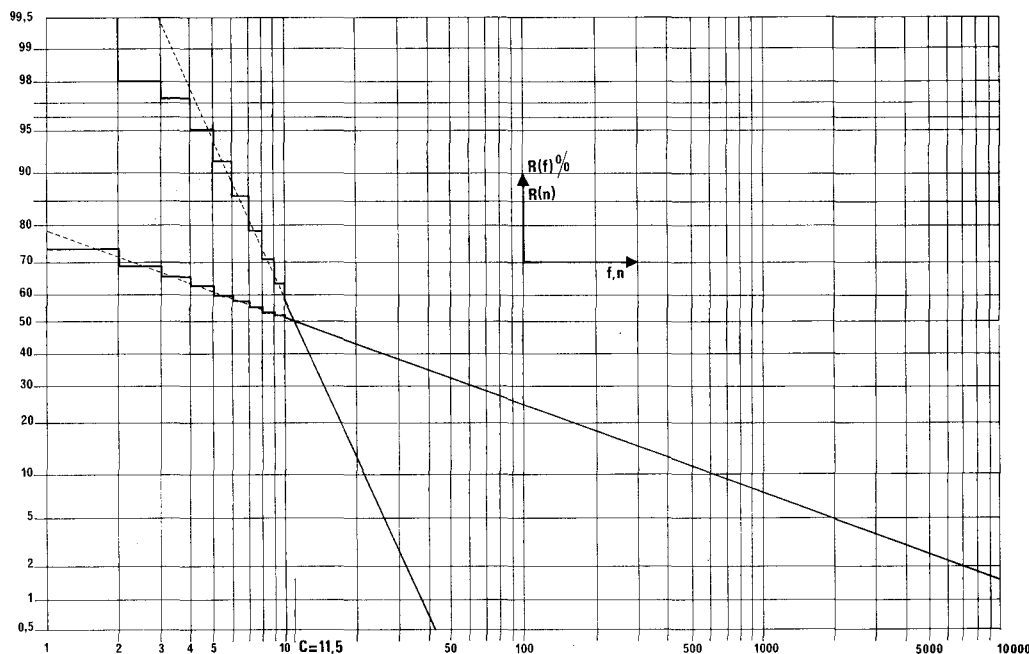


Fig. 3

The graph in figure 4 shows the two functions $R(f)$ and $R(n)$ on log-log paper, whilst the graph in figure 5 shows the two functions $p(n)$ and $p(f)$ which are the continuous approximation to respectively, the probabilities that n descriptors are posted to a document and that a descriptor has been used f times.

5 — DYNAMICS OF DOCUMENTATION SYSTEMS

In the preceding paragraphs we considered the descriptors and the documents as separate variables. This enabled us to demonstrate the symmetries that exist between the statistical distributions of f and n . In the dynamics study, however, we must forget that starting-point; for, obviously, the number of documents V is the independent variable in the system, whereas the number of descriptors (dependent variable) changes as the collection grows in volume.

In particular we may suppose that, if the collection is increased by dV , we must add to the dictionary $dN_1 = a_1 dV$ new descriptors to encode new concepts, whilst, on the other hand, dN_2 descriptors will have to be removed as no longer suitable. We can assume that dN_2 is

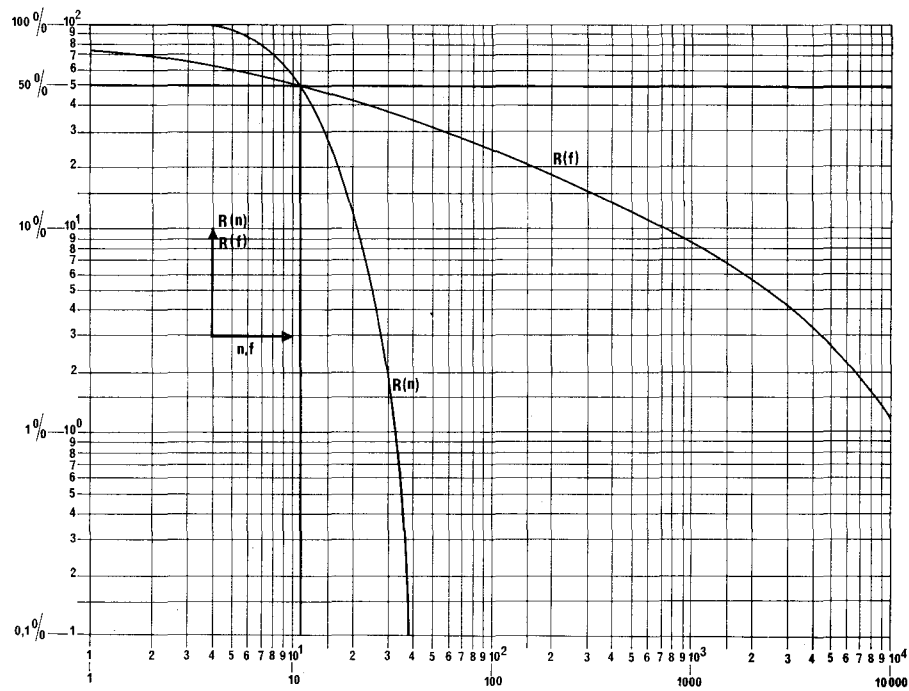


Fig. 4

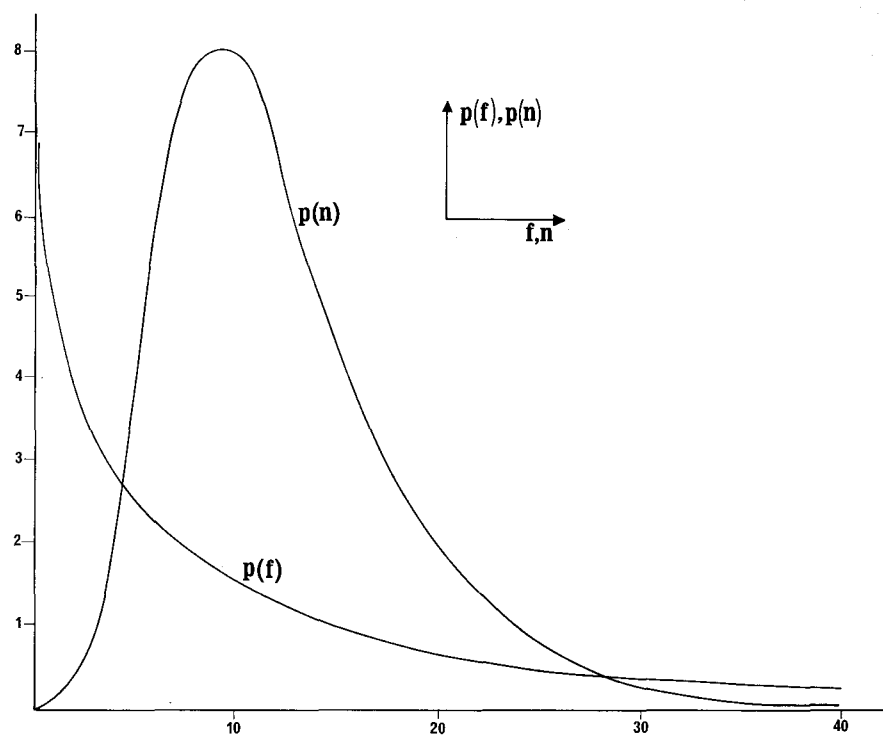


Fig. 5

proportional to dV and to the total number of dictionary descriptors, i.e., $dN_2 = -a_2 N dV$. Starting from this assumption, we find that the variation of the number of dictionary descriptors can be expressed by the relation

$$dN = a_1 dV - a_2 N dV$$

i.e., by means of the following linear, differential equation with constant coefficients. (See Note 3, page 24.)

$$\frac{dN}{dV} + a_2 N = a_1 \quad (29)$$

the solution of which is

$$N = \frac{a_1}{a_2} + K e^{-a_2 V} \quad (30)$$

in which K is dependent on the initial conditions.

As to the latter, supposing that we start with a ready-made dictionary containing a number of descriptors N_0 since, in this case,

$$\begin{aligned} \text{where } V &= 0 \\ N &= N_0 \end{aligned}$$

we have:

$$N_0 = \frac{a_1}{a_2} + K$$

that is,

$$K = N_0 - \frac{a_1}{a_2} \quad (31)$$

And if we impose the boundary condition we have, with $V \rightarrow \infty$, $N = N_\infty$, whence

$$N_\infty = \frac{a_1}{a_2} \quad (32)$$

Through (31) and (32), equation (30) becomes

$$N = N_\infty - (N_\infty - N_0) e^{-a_2 V} \quad (33)$$

an equation which in fact describes the law of growth of the number of dictionary descriptors as a function of the number of documents.

We can check (33) by the values in table IV obtained in respect of the Euratom CID documentation system.

TABLE IV

Order	N	V	Reference	Date
0	1 900	0	—	1962
1	6 470	100,000	EUR 500 . e I	1964
2	11 030	360,000	EUR 500 . e II	1966
3	12 070	642,000	present	1968

The values in this table were plotted on to the graph in figure 6 at the points 0, 1, 2, 3 ... As figure 6 shows, for the asymptotic value $N = 12,400$ the alignment is good. The value of a_2 is easily obtained from the same figure by finding the V for which

$$\frac{N_{\infty} - N_0}{e} = \frac{10\,500}{e} = 3\,880$$

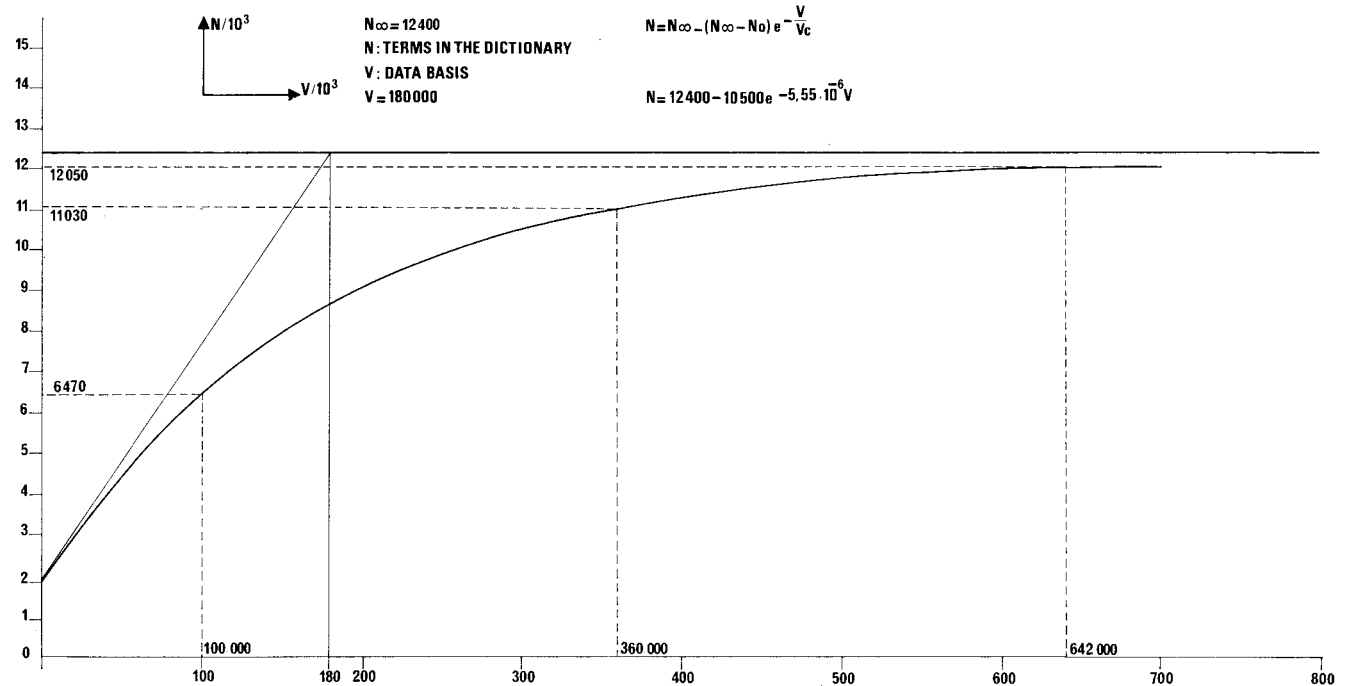


Fig. 6

This value is

$$V_c = 180,000$$

and represents the number of documents that would have to be indexed in order to attain N_{∞} if the law (33) had been linear.

For a_1 and a_2 we have

$$a_2 = 1/180.000 = 5,55 \cdot 10^{-6}$$

$$a_1 = N_{\infty} a_2 = 1,24 \cdot 10^4 \cdot 5,55 \cdot 10^{-6} = 6,9 \cdot 10^{-2}$$

Thus for the Euratom/CID System the law (33) becomes

$$N = 12.4000 - 10.5000 e^{-5,55 \cdot 10^{-6} V} \quad (34)$$

Let us consider the derivative of (33)

$$\frac{dN}{dV} = (N_{\infty} - N_0) a_2 e^{-a_2 V}$$

or

$$\frac{dN}{dV} = \frac{N_{\infty} - N_0}{V_c} e^{-V/V_c} \quad (35)$$

An analysis of (35), with "small" V/V_c values, leads to an interesting result. For by developing in series e^{-V/V_c} , we obtain

$$e^{-V/V_c} = \frac{1}{e^{V/V_c}} \approx \frac{1}{1+(V/V_c)}.$$

Through which (35) can be written as

$$\frac{dN}{dV} \approx \frac{N_\infty - N_0}{V_c} \cdot \frac{1}{1+(V/V_c)} = \frac{N_\infty - N_0}{V_c + V}$$

With small values of V/V_c expression (35) can be substituted by

$$\frac{dN}{dV} = \frac{10.500}{180.000 + V} \quad (36)$$

which is of the type

$$\frac{dN}{dV} = \frac{c_1}{c_2 + V} \quad (37)$$

an expression wholly similar to the one found by Houston and Wall.

NOTE 1

In the Euratom Nuclear Documentation System normalized term W_i used in indexing is called *DESCRIPTOR*. Any descriptor may be linked to others by relations (sub-ordination relations) like the following:

$$W_a \text{ USE } W_b .$$

In this case the descriptor W_b is automatically added when W_a is used. The descriptors which are not subordinated to others by a similar relation are called *KEYWORDS*.

The set $\{W_i\}$ of descriptors, i.e. of keyword and not-keyword standardized terms is called *DICTIONARY*. The list of all the descriptors of the dictionary completed by a list of forbidden terms (synonyms or homographic terms) is called *THESAURUS*. These forbidden terms should not be assigned at all; if they are, they are automatically replaced by the corresponding descriptor (in case of synonyms) or submitted to a documentalist for decision (in case of homographs).

NOTE 2

In some documentation systems, very general papers or review documents are split in *DOCUMENT UNITS*, each of which is dealing with analog matter. Such documentation systems may be considered as simple co-ordination systems if for "document" we mean each "document unit". This *SPLITTING* procedure is called, by some authors, a *PSEUDO-LINK* tool.

NOTE 3

The fact of considering a_1 as a constant means that the quantity of new concepts carried in the system by any new document analyzed is constant. This is rather true if the system does not change its *SCOPE*. In this case in fact any variation of N is due to the natural evolution of science.

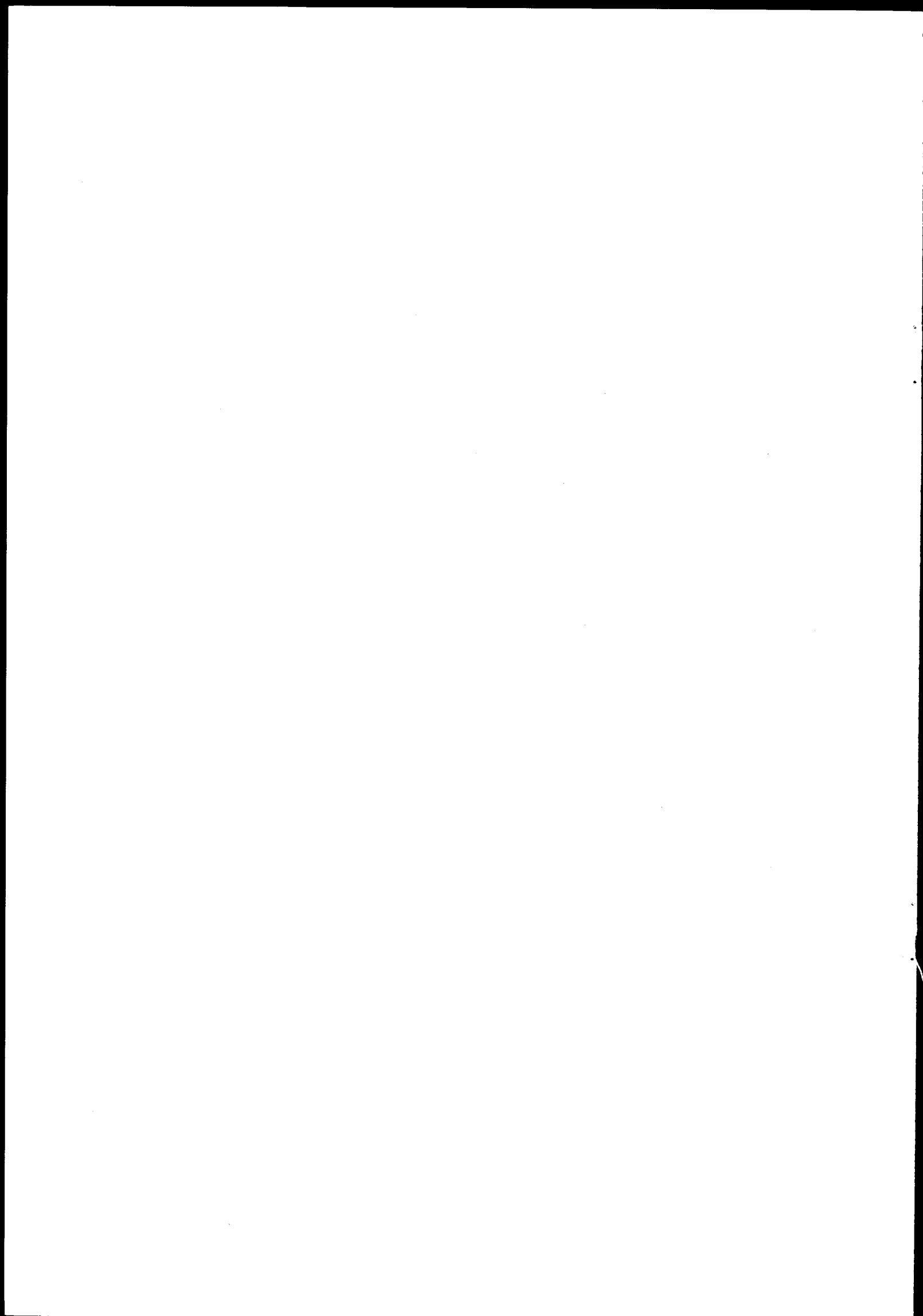
The consideration of a_2 as a constant means that the term-standardization procedure and the *INDEXING RULES* of the system are the same for the period considered.

Any change in the scope or in the indexing rules produces a transient in the curve of figure 6.

BIBLIOGRAPHY

- H. HEINZE, Wege zur Mathematischen Erfassung von Documentationsvorgängen. *Nachr. Dok.*, **18**, No. 3/4 (1967), pp. 100-103.
- H. HEINZE, Quantitative Analyse der optimalen Bedingungen für die Information aus der Fachliteratur in grösseren Unternehmen. *Nachr. Dok.*, **18**, No. 6 (1967), pp. 236-242.
- N. HOUSTON, and E. WALL, The Distribution of Term Usage in Manipulative Indexes. *American Documentation*, **15**, No. 2 (1964), pp. 105-144.
- F.H. LEIMKUNLER, The Bradford Distribution. *Journal of Documentation*, **23**, No. 3 (1967), pp. 197-207.
- L. ROLLING, and J. PIETTE, *Interaction of Economics and Automation in a large-size Retrieval System*. In Proc. FID/IFIP conf. Mechanized Information Storage, Retrieval and Dissemination, Rome, June 14-17, 1967.
- C.H. SCHULTZ, P.D. SCHWARTZ, and L. STEINBERG, A Comparison of Dictionary use within two Information Retrieval Systems. *American Documentation*, **12**, No. 4 (1961), pp. 247-253.
- P. ZUNDE, and V. SLAMECKA, Distribution of Indexing Terms for Maximum Efficiency of Information Transmission. *American Documentation*, **18**, No. 2 (1967), pp. 104-108.
- EURATOM - THESAURUS, EUR 500 e (1st edition, 1964).
- EURATOM - THESAURUS, EUR 500 e (2nd edition, part I, 1966; part II, 1967).





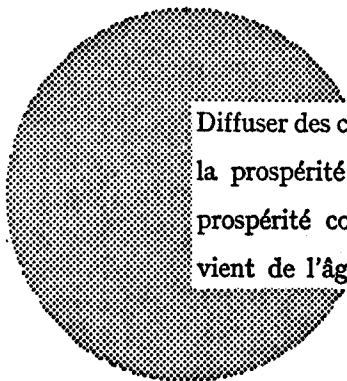
AVIS AU LECTEUR

Tous les rapports Euratom sont signalés, au fur et à mesure de leur publication, dans le périodique mensuel EURATOM INFORMATION, édité par le Centre d'information et de documentation (CID). Pour souscrire un abonnement (1 an : FF 75, FB 750) ou recevoir un numéro spécimen, prière d'écrire à :

**Handelsblatt GmbH
„Euratom Information“
Postfach 1102
D-4 Dusseldorf (Allemagne)**

ou à

**Office de vente des publications
des Communautés européennes
2, Place de Metz
Luxembourg**



Diffuser des connaissances c'est distribuer de la prospérité — j'entends la prospérité collective et non la richesse individuelle — et cette prospérité contribue largement à la disparition du mal qui nous vient de l'âge des ténèbres.

Alfred Nobel

BUREAUX DE VENTE

Tous les rapports Euratom sont vendus dans les bureaux suivants, aux prix indiqués au verso de la première page de couverture (lors de la commande, bien indiquer le numéro EUR et le titre du rapport, qui figurent sur la première page de couverture).

OFFICE CENTRAL DE VENTE DES PUBLICATIONS DES COMMUNAUTÉS EUROPÉENNES

2, place de Metz, Luxembourg (Compte chèque postal N° 191-90)

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

LUXEMBOURG
OFFICE CENTRAL DE VENTE
DES PUBLICATIONS DES
COMMUNAUTÉS EUROPÉENNES
9, rue Goethe - Luxembourg

DEUTSCHLAND
BUNDESANZEIGER
Postfach - Köln 1

NEDERLAND
STAATSDRUKKERIJ
Christoffel Plantijnstraat - Den Haag

FRANCE
SERVICE DE VENTE EN FRANCE
DES PUBLICATIONS DES
COMMUNAUTÉS EUROPÉENNES
26, rue Dasaix - Paris 15^e

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

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

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