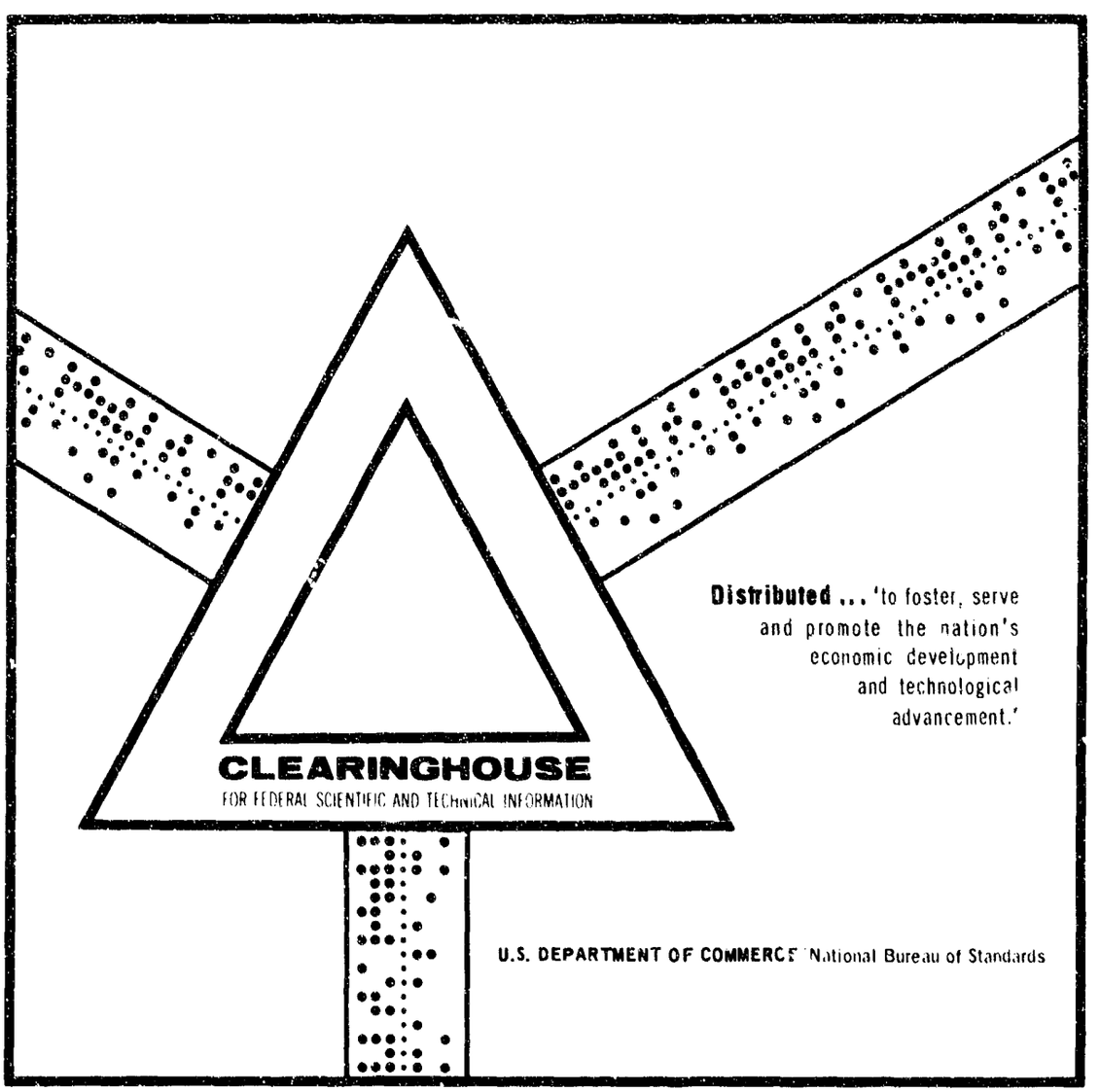
ONE OF THE METHODS OF ORGANIZING THE INFORMATION STORAGE UNIT OF A SYSTEM OF DATA RETRIEVAL AND PROCESSING (SPOD)

R. B. Askinazi, et al

Foreign Technology Division
Wright-Patterson Air Force Base, Ohio

30 September 1969



FTD-MT-24-196-69

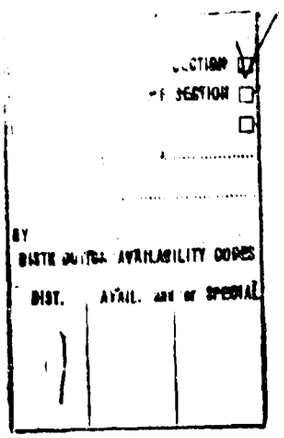
FOREIGN TECHNOLOGY DIVISION

ONE OF THE METHODS OF ORGANIZING THE INFORMATION STORAGE UNIT OF A SYSTEM OF DATA RETRIEVAL AND PROCESSING (SPOD)

by

R. B. Askinazi and I. L. Papina

This document is a machine translation of Russian text which has been processed by the AN/GSQ-16(XW-2) Machine Translator, owned and operated by the United States Air Force. The machine output has been post-edited to correct for major ambiguities of meaning, words missing from the machine's dictionary, and words out of the context of meaning. The sentence word order has been partially rearranged for readability. The content of this translation does not indicate editorial accuracy, nor does it indicate USAF approval or disapproval of the material translated.



EDITED MACHINE TRANSLATION

ONE OF THE METHODS OF ORGANIZING THE INFORMATION STORAGE UNIT OF A SYSTEM OF DATA RETRIEVAL AND PROCESSING (SPOD)

By: R. B. Askinazi and I. L. Papina

English pages: 12

Source: Nauchno-Tekhnicheskaya Informatsiya. Seriya 2. Informatsionnyye Protsessy i Sistemy (Scientific and Technical Information. Series 2. Information Processes and Systems), 1967, No. 12, pp. 37-39

UR/0447-67-000-012

PREPARED BY:

TRANSLATION DIVISION
FOREIGN TECHNOLOGY DIVISION
WP AFB, OHIO.

U. S. BOARD ON GEOGRAPHIC NAMES TRANSLITERATION SYSTEM

Block	Italic	Transliteration	Block	Italic	Transliteration
А а	*А а*	A, a	Р р	*Р р*	R, r
Б б	*Б б*	B, b	С с	*С с*	S, s
В в	*В в*	V, v	Т т	*Т т*	T, t
Г г	*Г г*	G, g	У у	*У у*	U, u
Д д	*Д д*	D, d	Ф ф	*Ф ф*	F, f
Е е	*Е е*	Ye, ye; E, e*	Х х	*Х х*	Kh, kh
Ж ж	*Ж ж*	Zh, zh	Ц ц	*Ц ц*	Ts, ts
З з	*З з*	Z, z	Ч ч	*Ч ч*	Ch, ch
И и	*И и*	I, i	Ш ш	*Ш ш*	Sh, sh
Й й	*Й й*	Y, y	Щ щ	*Щ щ*	Shch, shch
К к	*К к*	K, k	Ъ ъ	*Ъ ъ*	"
Л л	*Л л*	L, l	Ы ы	*Ы ы*	Y, y
М м	*М м*	M, m	Ь ь	*Ь ь*	'
Н н	*Н н*	N, n	Э э	*Э э*	E, e
О о	*О о*	O, o	Ю ю	*Ю ю*	Yu, yu
П п	*П п*	P, p	Я я	*Я я*	Ya, ya

* ye initially, after vowels, and after ъ, ь; e elsewhere. When written as ё in Russian, transliterate as yё or ё. The use of diacritical marks is preferred, but such marks may be omitted when expediency dictates.

ONE OF THE METHODS OF ORGANIZING THE INFORMATION STORAGE UNIT OF A SYSTEM OF DATA RETRIEVAL AND PROCESSING (SPOD)

R. B. Askinazi and I. L. Papina

Summary

This paper deals with one method of organizing the storage unit of a descriptor IPS (information retrieval system) of the SPOD type, the information array of which constitutes the totality of uni rm documents with ordered disposition of data within each of them. Three categories of data composing the retrieval form of a document have been defined classificational, base and key. For the organization of retrieval in a group of documents with a common series of similar data, use is recommended of the tableization method, the process of which can be automated by including a free, ordered table in the automatic dictionary of translation from the natural language to the IPYa (information retrieval language). Organization of the information storage unit anticipates the inclusion in each of its descriptors of a classificational criterion with corresponding subcriteria. The SPOD ensures output of data which corresponds to the level shown in the request, as well as data below this level, but pertaining to a higher section of the document.

An information-retrieval array can be represented by retrieval forms in the form of descriptor descriptions of the semantic contents of documents, using the tabular form for reflecting the quantitative

values of separate categories of data. It is expedient to define the three categories of data which compose its retrieval form:

- classificational;

- base;

- key.

Classificational data determine the position of a document in the storage system. They include bibliographic information and the storage address of the document in the information-retrieval array.

Base data designate objects, phenomena, or actions, the characteristics of which are the document. It is necessary to note that these data may not be included in the title of the document.

Key data are the necessary set of words from the text of the document, sufficiently accurate (for retrieval purposes) reflecting the details of its contents.

Like key data base data are the reflection of the contents of a document in its retrieval form. The separation of base data into a separate category is connected with the practice of request formulation. It is possible to assume that the category of data called "base" may appear only during the formulation of a request, since the consumer, for one reason or another, does not always know the details of the contents of a document. At the same time the form of the representation of these categories in the retrieval form of the document varies.

Let us examine, in a concrete example, the above proposed categories and the relationship of the separate data within them. Let us assume that the array contains reference data about various types of semiconductors, while the document - reference data about the P4A transistor [11].

The classificational data of the retrieval form of this document include bibliographic information and the document's address in the information-retrieval system utilized.

For the examined retrieval form of the document, the descriptor "P4A" belongs to the category of base data.

The contents of the document in the examined example have been divided into these sections: "assignment," "forming," "general data," "maximum permissible electrical data," etc. Each section is described by a group of key data. Here, some of these data, besides the purely semantic data, bear quantitative information. For example, in the section "general data" these concepts are used:

greatest height – 10 mm
greatest diameter – 31 mm, etc.

The character of the informational array permits determining the assumed character of the basic quantity of requests connected with the operation of retrieval: from a known series of defined data it is necessary to obtain the concrete address of the document and certain information contained in it in the form of a descriptor description and numerical values.

It is possible to assume the following basic types of requests to the IPS with the above array.

1. From the designation of the object, establish the values of its defined criteria.

2. From the defined criteria of the object, establish the values of other defined criteria.

3. From defined criteria, establish the designation of the object.

4. Check whether the concrete object, the designation or series of criteria of which are known, possesses other defined criteria.

5. Check at which values of the assigned criteria the indicated values of certain other criteria occur.

6. Issue documentary information (in the form of a concrete address of the document from the document) which possesses defined criteria [10].

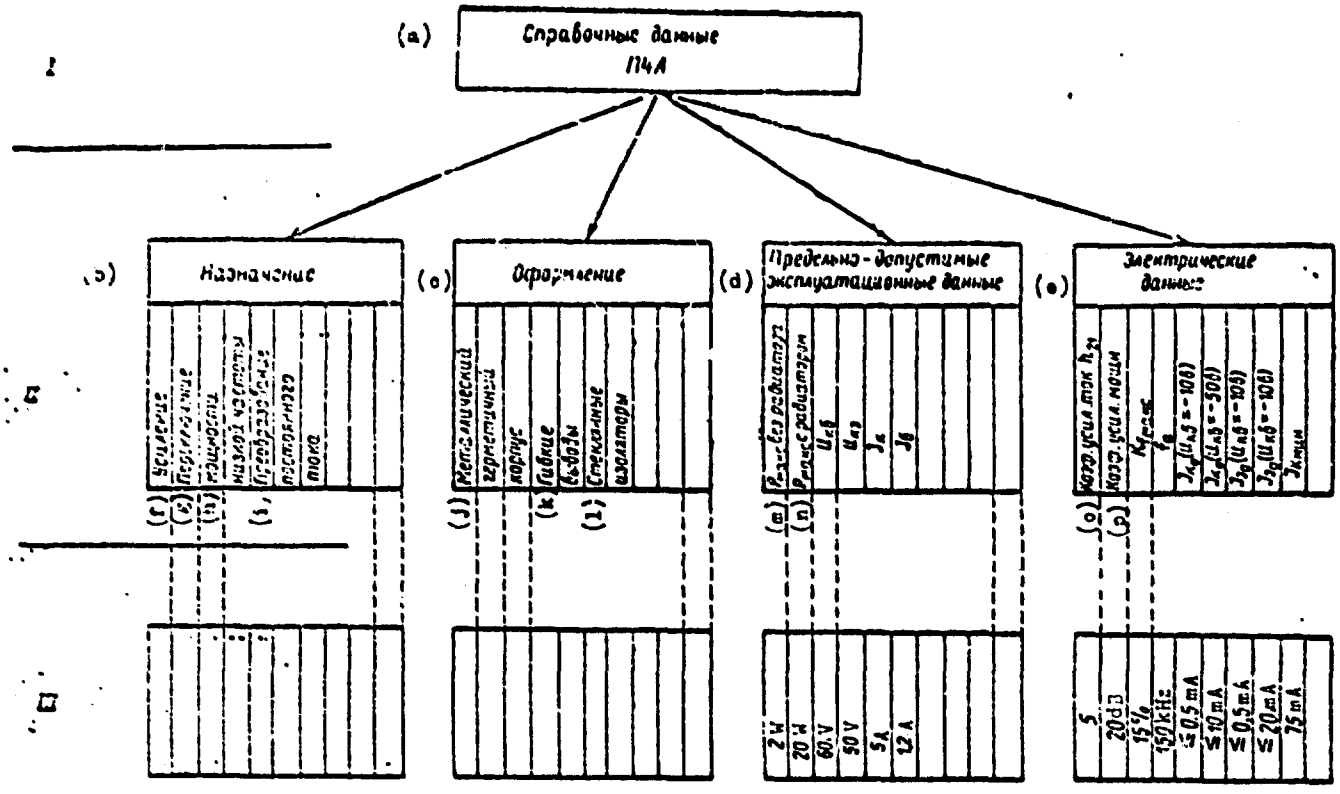
The realization of similar requests is connected with the fulfillment of a number of retrieval operations, which as, for example, sorting of data by a defined series of criteria, establishing the correspondence of the request to the document under the condition of entry (nonentry) of the defined series of request criteria in the retrieval form of the document, and so forth.

Two basic methods of realizing retrieval operations are known: direct and inverse. The specific character of the information material, containing a significant number of data with concrete quantitative values, and also the specific character of the requests, based on concrete quantitative values of defined data, render inverse search (in its pure form) practically unacceptable in view of the difficulty of realization. Therefore, for the examined type of information arrays, it is apparently more expedient to use the method of realization of retrieval operations. It is necessary to note the possibility of realization of retrieval with a combination of the direct and inverse systems. At first the inverse method determines the addresses of zones with retrieval forms corresponding to the document request; then by direct method, from these retrieval forms the necessary data are selected. The proposed variant is expedient when meeting two conditions:

1) the selection of retrieval forms of documents is made according to semantic values of the data without using their quantitative values,

2) the storage unit of retrieval forms of documents has random access.

The figure shows the structure of the retrieval form of a document selected as an example.



KEY: a) Reference data, P4A; b) Designation; c) Layout; d) Maximum achievable operational data; e) Electrical data; f) Amplification; g) Switching; h) Powers of lower frequency; i) Conversion of direct current; j) Metallic hermetically sealed housing; k) Flexible outputs; l) Glass insulators; m) $P_{макс}$ without a radiator; n) $P_{макс}$ with a radiator; o) Current amplification factor; p) Power amplification factor.

The descriptors which make up the retrieval form of the document are connected by sort-type relationships and are situated on three hierarchical levels. Requests to the system can be formulated on any of them. Here on level 1 the contents of the entire retrieval form should be issued; on level 2, the contents of the corresponding section, and on 3 – the value of separate documentary data.

One of the practical possibilities of realization of sort-type relations in the system is giving each descriptor of the retrieval form the criterion of the level. In this case, inside each retrieval

form all of the lower descriptors, characterizing the separate properties of the higher, will have the criterion level given to the higher.

The number of levels utilized in the system depends on the structure of the informational array and on the organization of the information storage unit: by-document or by-aspect. With by-document organization of the storage unit, for each document an individual retrieval form is composed which reflects all the aspects of its contents.

With by-aspect organization, for each aspect or group of aspects of a document its own retrieval form is composed. In this case, to each document several retrieval forms may correspond, each of which reflects only the separate aspects of its contents.

The selection of this or that organization of the storage unit is determined by the assumed character of the basic quantity of requests which, for an answer, require total examination of the document or an examination of only its individual aspects.

With direct organization of retrieval, the retrieval images of documents are recorded in the storage unit in the form of retrieval zones. Here each documentary category of data must be given an individual criterion. Argumentation of the necessity for alloting a criterion is revealed when examining the structure of the request.

Any request to the retrieval system contains:

1) a certain set of data for setting the correspondence of a certain set of documents to the request,

2) a certain set of criteria of data which are the answer to the request.

Thus, for an answer to the request "Give the brand of semiconductor triode (transistor) which is used for low frequency power

gain," it is necessary to rework the text of the request in such a way that instead of the words "bond," "semiconductor," and "triode," in the retrieval device a criterion was introduced which corresponded to the collection of these words. This reworking of the request test can either be done manually, with the aid of translation tables, or automatically with a special dictionary.

It should be noted that with such an organization of the retrieval zone, the answer includes all of the descriptors of the retrieval form noted by the criterion indicated in the request. Thus, if the bibliographic information and the address of the document are given in the retrieval zone under one criterion, "classification data," then to the request "Indicate the concrete address of the document" the system given an answer in which, besides the address, bibliographic information also will be shown — i.e., in the answer noise will appear.

To lower the noise emitted by the system, it is necessary to detail the criteria of the accompanying data.

One of the methods of organizing documentary data in the retrieval zone is tableization — representing these data in form in which the semantic meaning of each is determined by its position in the ordered list. Tableization should be used if there is a sufficiently large group of documents with a common series of similar (in meaning) data, which has various quantitative value.

In tableization, in the retrieval zone of the storage unit the numerical values and reference numbers of corresponding data are filed from the combined ordered table of their semantic meanings. The process of data tableization when compiling retrieval zones can be automated by including a combined ordered table in the automatic translation dictionary from the natural language to the information retrieval language (IPYa). Of course, tableized data in the retrieval zone must have the criterion "table."

Fundamentally it is possible to tableize purely semantic data of the descriptor description of documents. In this case, the corresponding words of the retrieval zone would have double information. However, a semantic table in comparison with an ordinary descriptor description gives no essential advantage, since the number of words in the retrieval zone does not change, and the volume of the combined list of tableized data increases.

The essential question of organizing the information storage unit of an SPOD is the translation of words of the natural language of documents into IPYa. Here we must consider that the translation method must be useful both for the text of the document as well as for the text of requests. This translation can be carried out manually or automatically. Furthermore, during translation into the IPYa additional classificational operations can be carried out. Here the texts are given additional organization by means of preliminary processing, leading to indexing.

In the simplest case the IPYa of a system can be natural language translated into machine form. Here, to decrease the word length, different artificial methods of compressing word codes [3] are used.

An inherent difficiency of this method is the need for the consumer to use, when compiling the request, the same words of the natural language which were used to compile the retrieval form, under the condition that this set of words utilized in the retrieval form is unknown to the consumer. Use of this method is expedient in systems with a formalized setting [Translators Note: This is the literal translation of the word "usroyavshiysya." No appropriate translation found in available sources.] language.

In highly organized systems, dictionaries are used. Here the words of the document and the request are translated into the IPYa in dictionary terms, which eliminates losses during retrieval due to ambiguity of words utilized.

Furthermore, the dictionary can realize basis-type and textual relationships; it can also give additional organization to the set of words of the retrieval and request forms.

The use of textual relationships in the SPOD is apparently inexpedient, since according to the results of experimental investigations, the positive effect obtained with their use does not justify the necessary complications of the system for their realization [2].

Moreover, it is possible to assume that the specific character of the information array of the SPOD, other things being equal, ensures an additional drop of information noise when answering a request.

The dictionary ensures translation of the words of the natural language into IPYa by means of consecutive comparison of input words with the total volume of words used in the dictionary.

The problem of synonymy is solved by including in the dictionary a group of synonyms and conditionally equivalent words.

The problem of homonymy and polysemy is solved by combining the word of the homonym with a group of words which explain its semantic meaning.

The problem of calculating base-type relationships is solved by introducing classes of classificational criteria into the descriptor.

Classificational criteria determine the hierarchic belonging and sort-type relationships of data of the information array.

The classificational criterion joins a series of subcriteria:

1) of the belonging of data to one of three categories: classificational, base, or key;

2) of the designation of divisions of a document which are at one level (in the considered example, there are eleven such divisions: "general data," "assignment," "sharping," "conditions of exploitation," "electrical parameters," etc);

3) of the level, if divisions of the document are disposed at different levels;

4) of the table for indicating the entry of corresponding data in a class of tableized data;

5) of the quantity of words in a word combination, translating into IPYa by one descriptor (for example, the word combination "maximum permissible electrical data" is translated into IPYa by one descriptor);

6) of the group of synonyms or conditionally equivalent words;

7) of a word from a group of synonyms, subject to print out when using the dictionary in the mode of translation from IPYa to the natural language, in the mode of output of results.

In the dictionary digits which are allocated for recording the translation of a word of a language into IPYa, either the serial numbers of the tableized data which are the translation of the semantic meaning of this data into the IPYa are stored. The quantitative values of tableized data are not recorded in the dictionary.

During the formation of the classificational criterion it is possible to manage without the subcriterion "table" during the following organization of tableization. The serial numbers used as descriptors during translation to IPYa of semantic data are limited, for example, from below by a specified number "a." In this case, tabular data are translated by descriptors, the values of which are disposed in the form of serial numbers up to this number. The

belonging of a word to the table is determined in this case by a special analyzer by the criteron of the value of the descriptor, lying within the limits of from one to "a."

The above dictionary organization places corresponding conditions on the organization of the storage unit. Each descriptor of the storage unit has a classificational criterion with corresponding subcriteria. In the digits allocated for recording of the value of the descriptor either the serial numbers of the descriptors in the dictionary are recorded, or, in the case of tableization, the numerical values of tableized data are recorded.

The semantic meanings of tableized data are recorded in the digits of the subcriterion "table" in the form of the serial number of the semantic meaning of this data in the ordered tableized list.

As was mentioned above, the request includes a group of words for a description of the criterion of data which are subject to being output as an answer.

This group of words, including synonyms and conditionally equivalent words, enters into the supplemental dictionary list. In distinction from the basic composition of the dictionary, the words of the supplemental list are translated into IPYa only by classificational criterion. Since one and the same word can enter into both groups of the dictionary list (basic and supplmental and, consequently, they have to be translated into IPYa differently, it is necessary to introduce a criterion of the group of dictionary composition. The belonging of a word to this or that group is established during input of the request by singling out that part of it which pertains to the description of the criterion of issued data. In the information storage unit this criterion is absent.

When using the proposed method of organization of the information storage unit, the SPOD ensures output of data corresponding to the level shown in the request, and also lying below this level, but pertaining to a higher section of the document.

Bibliography

1. Mikhaylov A. I., Chernyy A. I., Gilyarevskiy R. S. Osnovy nauchnoy informatsii (Bases of scientific information). M., Izd-vo "Nauka", 1965, 654 str.

2. Sokolov A. V. Poteri informatsii i informatsionnyy shum v deskriptornykh IPS i v kartoteke, sistematizirovannoy po UDK. (Po rezul'tatam eksperimental'nykh issledovaniy) (Losses of information and information noise in a descriptor IPS and in a card file, systematized by UDC (According to the results of experimental investigations)). "NTI", 1964, No. 9, 20-23.

3. Rayli V. (W. B. Riley). Ispol'zovaniye elektronnykh vychislitel'nykh mashin v informatsionno-poiskovykh sistemakh (Use of electronic computers in information retrieval systems). "Elektronika", 1966, No. 5, 63-66.

4. Bernshteyn E. S., Lakhuti D. G., Chernyavskiy V. S. Voprosy teorii poiskovykh sistem (Questions on the theory of retrieval systems). M., VNIIEM, 1966, 132 str.

5. Bernshteyn E. S., Lakhuti D. G., Chernyavskiy V. S. Nekotoryye voprosy postroyeniya deskriptornykh IPS (Some questions of construction of descriptor IPS). "NTI", 1963, No. 1, 31-33.

6. Bernshteyn E. S. K voprosu ob avtomaticheskom indeksirovanii (Concerning the question of automatic indexing). "NTI", 1963, No. 10, 22-25.

7. Sokolov A. V. Issledovaniye poter' informatsii i informatsionnogo shuma v deskriptornykh IPS (Investigating information losses and information noise in descriptor IPS). "NTI", 1965, No. 12, 23-28.

8. Simmons R. Avtomaticheskoye polucheniye otvetov na voprosy, sformulirovannyye na angliyskom yazyke (Automatic production of answers to questions formulated in the English language). "NTI", 1966, No. 4, 27-31; No. 5, 18-26.

9. Plyatsidevskiy I. A. Informatsionnyye sistemy v tekhnike i ekonomike (Information systems in technology and economics). M., "Moskovskiy rabochiy", 1966, 175 str.

10. "Tranzistory i poluprovodnikovyye diody" ("Transistors and semiconductor diodes") edited by I. F. Nikolayevskogo. M., "Svyaz'izdat", 1963, 646 str.

DATA HANDLING PAGE

01-ACCESSION NO.	02-DOCUMENT LOC	03-TOPIC TAGS
TP9501458		information storage and retrieval, information processing, computer language, computer coding

04-TITLE ONE OF THE METHODS OF ORGANIZING THE INFORMATION STORAGE UNIT OF A SYSTEM OF DATA RETRIEVAL AND PROCESSING (SPOD)

05-SUBJECT AREA

09

06-AUTHOR CO-AUTHORS	10-DATE OF INFO
ASKINAZI, R. B. ; 16-PAPINA, T. L.	-----67
11-SOURCE NAUCHNO-TEKHNICHESKAYA INFORMATSIYA. SERIYA 2. INFORMATSIONNYYE PROTSESSY I SISTEMY (RUSSIAN)	12-DOCUMENT NO. FTD-MT-24-196-69
	69-PROJECT NO. 6050202

63-SECURITY AND DOWNGRADING INFORMATION	64-CONTROL MARKINGS	97-HEADER CLASN
UNCL, 0	NONE	UNCL

76-REEL FRAME NO.	77-SUPERSEDES	78-CHANGES	40-GEOGRAPHICAL AREA	NO OF PAGES
1890 0291			UR	12
CONTRACT NO.	**X REF ACC. NO.**	**PUBLISHING DATE**	**TYPE PRODUCT**	**REVISION FREQ**
	65-	94-	TRANSLATION	NONE

STEP NO.	ACCESSION NO.
02-UR/0447/67/000/012/0037/0039	

ABSTRACT

(U) This paper deals with one method of organizing the storage unit of a descriptor IPS (information retrieval system) of the SPOD type, the information array of which constitutes the totality of uniform documents with ordered disposition of data within each of them. Three categories of data composing the retrieval form of a document have been defined classificational, base and key. For the organization of retrieval in a group of documents with a common series of similar data, use is recommended of the tabulation method, the process of which can be automated by including a free, ordered table in the automatic dictionary of translation from the natural language to the IPYa (information retrieval language). Organization of the information storage unit anticipates the inclusion in each of its descriptors of a classificational criterion with corresponding subcriteria. The SPOD ensures output of data which corresponds to the level shown in the request, as well as data below this level, but pertaining to a higher section of the document.