Compilation of a Thesaurus and Total Index for *Nihon Kagaku-Gijutsu-Shi Taikei* by Means of a Computer

Tetsuo TOMITA* and Kazutoshi HATTORI**

I. Introduction

This paper presents a report on the compilation of a thesaurus and total index of history of science and technology in Japan. Using an electronic computer we prepared this from the *Nihon Kagaku-Gijutsu-Shi Taikei* (*History of Science and Technology in Japan*) 25 volume were edited by the History of Science Society of Japan and published from *DAI-ICHI HOKI SHUPPAN* Co. Ltd. between March 1964 and August 1970. In March 1968, when publication of the Nihon Kagaku-Gijutsu-shi Taikei (hereafter referred to as *Taikei*) was nearing completion, Professor Mitsutomo Yuasa proposed the preparation of a thesaurus of the history of science and technology in the form of a total index covering all 25 volumes to be published as a supplementary volume. Having consulted with members of the History of Science Society of Japan, he assigned us to prepare the total index in Cross Index style. After analysing all the documents in the *Taikei* from March to July, 1971, we completed the work in December 1971. The supplementary volume was published in January, 1972. Nihon Kagaku-Gijutsu Kenshujo (Institute of Computation Centre, Union of Japanese Science and Engineering; hereafter referred to as JUSE) voluntarily collaborated with us in the programming and running of an electronic computer.

Definition of words

In this paper we use the following definitions and abbreviations:

"Vol. X" is the Xth volume.

"n vols." is number of volumes.

"DN" (document number) refers to a document number as given in *Taikei*.

"IN" (indexing number) is used for a re-ordered document number as given in the thesaurus.

"KW" means Key-word.

"KWN" (Key-word number) refers to a specified number (or address number) of a Key-word used for the computer.

---

* The Patent Office under The Ministry of International Trades and Industries.

** The National Diet Library.

"Word" is used for a natural word.
"Term" is a word used as a classification.
"Key-word" is a word used as a term of the thesaurus for indexing.

II. General Principles

When Key-words are arranged and packetted in a classification system, resemblant or roughly similar words appear together, thus permiting the use of a hierarchical order in the classification of cross references in a thesaurus. A multi-dimensional classification system is one in which a certain Key-word occurs in various classifications. One merit of using a multi-dimensional classification system in compiling a thesaurus is that it enables one to link coupling facets of Key-words in independent fields such as science, industry, policy and economy. On the other hand, a catch word index, a Key-word list arranged in the order of the Japanese syllabary, results in sets of words of similar pronounciation, such as "industrial educations" and "industrial psychology." These sets permit a convenient check of related terms. In compiling the thesaurus of history and technology in Japan, we used a combination of the multi-dimensional classification and the catch-word index. In choosing the format of the thesaurus, we adopted the "Cross Index" style used by the publishers of Biological Abstracts, since it is most suitable for an index which is printed and bound as a book. Index numbers of Key-words are printed within the Key-word list in the order of Japanese syllabary.

III. Special Characteristics of a Thesaurus of Historical Documents

Although the thesaurus has often been used in assenbling information in the fields of science, technology, management and economy, no previous attempt has been made, to the best of our knowledge, to prepare a thesaurus of historical documents. Let us consider some characteristics of historical documents. Historical documents are similar to the experimental or the observable data used in pure sciences. They differ from scientific papers or news items. In that the data which they contain must be processed before it yields useful information. The characteristics of the historical documents include:

(1) The contents of documents are limited to special events or special circumstances in particular areas and, therefore, they have no universal validity.
(2) As each document is independent and isolated from other documents, one document cannot be used in place of another. This characteristic rejects the making of abstracts of historical documents.
(3) Some historical documents are valuable not because of their contents but because of the circumstances under which they were made or the influences they exerted on society.

(4) Documents relating to any social problems may be historical documents of science and technology.

(5) The form, style, language and other bibliographical data of documents

Table 1.

Volume	Title	No. of pages[1]	No. of chapters	No. of documents (A)[2]	No. of documents (B)[3]	No. of words in index ($c = d + f$)	No. of personal names (d)	No. of foreign names (f)	c/b	c/d	f/d
1	Outline history I	582	14	230	257	829	316	29	3.24	0.381	0.092
2	Outline history II	596	15	130	217	887	329	53	4.09	0.370	0.161
3	Outline history III	550	11	197	213	1,359	301	20	6.37	0.222	0.065
4	Outline history IV	532	10	186	229	1,179	116	—	5.14	0.098	0
5	Outline history V	558	13	201	218	1,062	119	19	4.87	0.112	0.160
6	Philosophy	552	10	145	187	758	293	43	4.05	0.386	0.150
7	International	566	14	139	152	1,794	558	464	11.82	0.311	0.800
8	Education I	562	14	160	259	1,405	407	94	5.42	0.290	0.231
9	Education II	542	16	170	205	1,521	509	58	7.41	0.335	0.114
10	Education III	578	14	168	207	1,779	270	20	8.60	0.152	0.739
11	Natural environment	594	17	199	229	1,834	681	41	8.00	0.371	0.601
12	Mathematical science	598	16	195	263	1,385	407	85	5.26	0.294	0.208
13	Physical Sciences	576	12	164	198	1,588	453	79	8.00	0.286	0.174
14	Astoronomy & earth science	614	15	288	349	1,524	612	87	4.37	0.402	0.142
15	Biological Science	538	10	172	194	1,073	404	121	5.49	0.376	0.300
16	Civil engineering	542	10	162	173	1,377	207	27	7.95	0.151	0.130
17	Architecture	590	12	182	275	533	199	50	1.94	0.374	0.251
18	Mechanical engineering	580	19	191	208	1,380	296	36	6.63	0.214	0.123
19	Electrical engineering	544	14	171	240	1,110	275	51	4.63	0.248	0.185
20	Mining & metallurgy	554	12	188	211	1,392	493	71	6.59	0.354	0.144
21	Chemical engineering	542	10	203	243	1,051	481	16	4.33	0.457	0.033
22	Agriculture I	572	11	176	231	1,783	251	22	7.71	0.141	0.087
23	Agriculture II	536	10	130	149	1,544	165	10	10.37	0.107	0.061
24	Medecine I	540	10	131	153	1,013	291	61	6.67	0.254	0.210
25	Medecine II	570	9	173	190	1,083	222	27	7.86	0.205	0.122
Total		14,088	318	4,451	5,450	32,243	8,655	1,584			
Average		564	12.7	178	218	1.290	347	63.4	5.92	0.269	0.183

(1) Number of pages does not include preface pages etc. because irrelevant for select a index-word in context.

(2) It is counted from simple documents.

(3) Documents No. 3-5A, 3-5B...etc. were counted separately.

may be important to historians, as for example, the first editions of text books in national languages or architectural illustrations.

These characteristics of historical documents have a bearing on thesauri as follows:

(A) Proper nouns such as names of persons, institutes, universities, companies, congresses, acts, and treaties are necessary Key-words.

(B) Famous notorious events, *e.g.*, large earthquakes and great fires, or social tendencies, *e.g.*, "Pax Britannica," are important Key-words.

(C) Chronological and geographical data are necessary.

(D) The form or style of documents, *e.g.* language, style of printing, are relevant in many documents such as diaries, essays, autobiographies, official documents, statistical data, graphs, maps, pictures, and blueprints.

(E) Terms in all fields are necessary because the various fields in which historical documents are related to developments in science and technology are unlimited.

IV. On Taikei and Indexing Numbers

"Taikei" consists of 25 volumes of source materials in the history of Japanese

Table 2. No. of documents per chapter in 21 volumes.

2-a

DN	No. of materials	DN	No. of materials	DN	No. of materials	DN	No. of materials
1	210	11	184	21	45	31	4
2	269	12	161	22	35	32	2
3	269	13	142	23	29	33	2
4	267	14	122	24	23	34	2
5	267	15	104	25	19	35	2
6	261	16	92	26	14	36	2
7	253	17	79	27	13	37	2
8	237	18	66	28	12	38	2
9	221	19	56	29	8	39	1
10	203	20	49	30	8	40	1

2-b

End number of DN	No. of materials	End number of DN	No. of materials
1	503	6	369
2	461	7	347
3	442	8	317
4	416	9	286
5	392	0	261

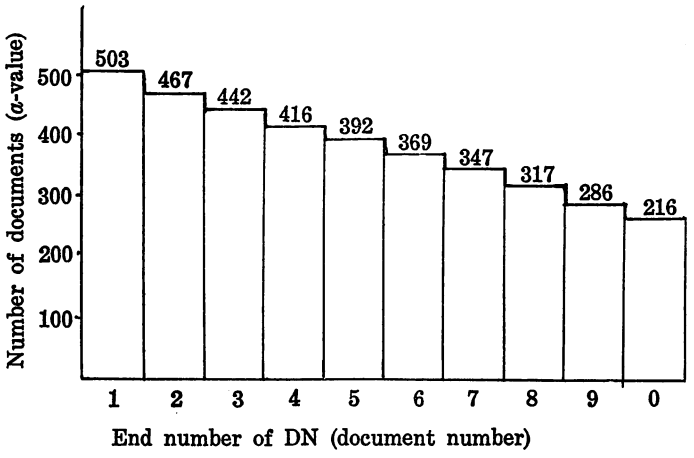

Fig. 1. End number of DNs.

science and technology during the last 100 years. Each volume consists of min. 9, max. 19, average 12.7 chapters, and contains min. 130, max. 288, average 178 selected documents. The total number of documents is 4,451 and the average number of documents per chapter is 14.1 (A-value documents in Table 1).

Each document has a document number which refers to be a chapter and to the order within each chapter, for example, 3–4, 3–5, 3–6. Some document numbers are divided further, as 5–4A, 5–4B and so on. When we count documents, taking into account these subdivisions, they total 5450 (min. 149, max. 349, average 218 documents per volume, 15.4 documents per chapter).

When these document numbers are used in the Cross Index, the numbers ending with 1 (for example 3–1, 5–11, 6–21) are nearly double (1,93 times) of those ending with 0 (2–10, 3–20, 6–30) (in Table 2a-b, Fig. 1), the length of the column ending with 1 needs twice that of the column ending with 0. Therefore, another serial, "indexing number," is used instead of the document number to refer to each B-value document. Thus, "Vol. 1 document 1–1" is indicated as "1001," and "Vol. 19. document 2–14C" is "6037."

V. Index of 25 Volumes and Key-Words

Generally, in preparing a thesaurus, analysis of all documents is seldom achieved since, in many cases, the number of documents to be analysed increases with time, resulting in a large accumulation in future. Thus, usually, one prepares a thesaurus with old documents supplementing it with new terms later. In our case, however, all documents for analysis have either been printed or are being prepared for printing. Therefore there will be no future accumulation of documents. Moreover, there will be no opportunity to maintain the thesaurus after publication. In addition, no arbitrary or random characteristic,

which would indicate statistically currents and trends of history, can be ascertained from *Taikei* since it was compiled from selected or typical documents only. Therefore, the method of sample analyses of documents was hardly appropriated. This pointed to the need to analyse all documents, a task requiring the investigation of 14,088 pages of documents. Therefore, we considered using the index included at the end of every volume as a rough indication of the content of the documents in the volume. The number of terms appearing in each index is given in Table 1.

Each index contains min. 533, max. 1834 average 1,290 words. These are regarded as valuable guides in compiling the thesaurus owing to the following factors:

(1) Words given in the index at the end of every volume are natural words.
(2) It is certain that those words exist in some parts of the volume.
(3) It is supposed the editors of each volume thought those words somewhat important.
(4) As every volume has an index at the end, we can see which words are important in the 25 volumes as a whole without twice analysing the documents.
(5) Each Index has deviation and bias of picking up words, but we can cancel such deviation by amassing the words of the 25 indices, since the diviation or absence of common words of a certain volume may be corrected by other volumes. If words are not supplied, their absence justly indicates the lack of documents in a particular field. It is unnecessary to supply such words in our work.

VI. Index of Every Volume

The index at the end of every volume contains min. 533, max. 1,834,

Table 3. List of frequency of terms used in index attached to end of Volume 1.

No. of pages corresponding to each index words (A)	No. of index words (B)
1	686
2	98
3	30
4	6
5	7
6	1
7	1
Total	829

Note: $(\sum A_i \cdot B_i)/(\sum B_i) = 1.26$.

average 1,290 words, the total for the 25 volumes is 32,243 words. This means that 5.92 words have been extracted from a B-value document or 2.28 words from a page. But some words in an index may correspond to several pages, so we have defined the frequence of words in the index of Volume 1 as referring to the number of pages in which the words occur (see Table 3). By this count, a word is extracted from 1.26 places. Therefore, the number of words extracted from all 25 volumes is 40,500, or 2.88 words per page, 7.45 words per document. We concluded that the words of these indices have suitable depth as resources of Key-words.

VII. Names of Persons

The words of indices at the end of every volume were written on cards and then the cards were classified. In classifying the cards, all cards of personal names were first extracted. The number of personal names in an index is min. 116, max. 681 and total 8,655; amongst these, the number of foreign names is min. 0, max. 461, average 63.4 and total 1,584. At first, terms were selected from personal names according to frequency,[1] resulting in 824 names. By adding names of persons who appear more than once as writers of documents, a total of 1,137 names were obtained as Key-words.

VIII. Analysis of Words in Indices and Extraction of Key-Words

The total number of words in the index of all volumes is 32,243. After extracting the 8,655 personal names, 23,588 words relating to subjects remained. These were classified and necessary terms were extracted.

The thesaurus was compiled in the following five stages.

Stage 1 Rough sorting of cards in several (first 10 and finally 40) fields.

Stage 2 Detailed re-sorting of Stage 1 cards by dividing each field into from 10 to 40 groups (a group contains from 10 to 40 words).

Stage 3 Writing up words on cards in a group.

Stage 4 Picking up or extracting important words from written up words.

Stage 5 Preparation of the thesaurus.

- 5A Making small hierarchical term lists with words extracted from several groups.
- 5B Sorting the term lists into seven major fields, *i.e.*, Research and Science; Technology and Industry; Policy and Economy; Social and Labour Problems; Life-Medical and Educational Problems; Philosophy, Religion and Fine Arts; and Unusual Events.
- 5C Fusing small classifications into one field.

[1] T. Tomita and K. Hattori "History of Science Society (ed.): Nihon Kagaku-Gijutsu-shi Taikei (History of Science and Technology in Japan), 25 Vols., 1864–1970." (Japanese Studies in the History of Science No. 9, 1970).

5D Fusing all classifications into the seven fields, making uniform different words of similar meanings and giving them cross references.

5E Editing a term list in the order of the Japanese syllabary.

Attention was also paid to following four points.

(1) When we sorted cards in Stage 1, we took vertically crossing divisions of fields, in contrast to the divisions of the 25 volumes. We also took various vertically crossing facets in classifying words during the assortments.

(2) Currents or trends of developments of society, *i.e.*, historical trends, are often changed by unpredicted or accidental events, such as wars, revolutions, earthquakes, typhoons and floods, and epidemic diseases. In China and Korea, historians paid attention to such unusual events as premonitory symptoms of social perplexity. Therefore as will as myths we devised a special facet entitled "Unusual events and phenomena."

(3) Many kinds of proper nouns are necessary. Besides personal names, we took the following proper nouns as facets: academic societies, research institutes, research organizations, international organizations, congresses, governmental organizations, companies, societies, labour unions, acts, treaties, objects of policies and social movements, social tendencies, universities, schools, libraries, museums, hospitals, important edifices, journals and newspapers.

(4) To construct the thesaurus, we combined different facet to construct a field or group, for instance, for the classification of contents of researches, we combined the three facets "branches of learning," "concepts and objects of research," and "experimental techniques."

Further, our thesaurus has the following distinctive points:

(5) One part of the classification system was based on names of places such as continents, oceans, seas, countries, prefectures, important cities, mountains, rivers and lakes. Since borders of nations change because of war, we used the borders of 1940 as indicators of Key-words.

(6) In chronological data, we took three different sets of Key-words: Key-words indicating era, those indicating each year since 1800 A.D. and those indicating five-year intervals.

(7) We prepared a facet for styles or kinds of documents. Key-words of the facet are also included in other facets, for instance, "abstracts of papers" is also classified under "tools of sciences" and "publications."

IX. Structure of the Thesaurus

Finally, we obtained three term-lists: a classified Key-word list, a Key-word

list in the order of the Japanese syllabary and a list of personal names in the order of the Japanese syllabary.

The classified Key-word list consists of general terms classified into the aforesaid seven fields; place names grouped in classifications such as seas, countries, prefectures, important cities, mountains, lakes, and rivers: and chronological divisions, such as eras, five-year intervals and individual years. Key-words are grouped and serial numbers are given to each Key-word to indicate addresses in the classification. Similar numbers are also given to Key-words in the list following the order of the Japanese syllabary. Therefore, a Key-word in order of Japanese syllabary produces a kind of cross reference by relating different Key-words from the classification systems. As some of Key-words appear at several different places in the classified list, we can make multiple relations between Key-words in different classifications.

A part of classified list is shown in Fig. 9-1, and a list following the order of Japanese syllabary in Fig. 9-2.

In the right columns of both lists are Key-word numbers for input into the computor.

As the Key-words consist not of sentences, but of single or compound nouns, hierarchical relations between words are not strictly logical. An item in a hierarchy is not a word including all terms, but a noteworthy or representative word. Sometimes we took plural terms as headings of groups or items.

Terms having hierarchical relations are regarded as Broader Terms or Narrower Terms, and terms of same rank are regarded as Related Terms.

The serial numbers indicating adresses in the classification have no function except that of indexing Key-words. After finding a Key-word in the list in the order of the Japanese syllabary, it is easy to look for the word in the classified list with the aid of the serial numbers. (regional numbers of classification in Fig. 9)

The following symbols are used as cross references.

➡ See.

⇨ See also.

⬅ Used for.

卐 Cross both words.

( ) Words not used as headings of the index.

Symbols are used, for example, as follows:

(1) Rivers (⇨卐 800) [800 is the address of names of places.]

(2) (Electric irrigations ➡ Electricity 卐 Irrigations).

Among the symbols, we think "Cross both words" is unique and very useful in the deletion of rarely used words. The symbol ( ) is sometimes used with "See," and plural terms of similar meanings, for example, "phonographs" and

"gramophones," "valves" and "vacuum tubes," can be reduced to single terms.

In the Key-word list which follows the order of the Japanese syllabary, words of similar pronounciations are arranged together, thus giving "Related Terms" of a different type.

Therefore, this thesaurus is very rich in associations of words, providing hierarchical relations, terms of the same rank, cross reference symbols and lists in the order of the Japanese syllabary. It contains 3,709 subject descriptors, 1,137 personal names, 200 chronological terms and a total of about 5,000 words.

X. Number of Pages of the Index

As printed matter, the index must be organized into a limited number of pages. Each term of the index consists of the head line (Key-word) and 10 columns of Indexing Numbers. Indexing Numbers are divided into 10 groups according to the final number and are arranged in corresponding columns. The number of pages of the index is calculated as the product of the number of Key-words and the average lines for a Key-word.

When an Indexing Number corresponds to a Key-word, a certain column is filled with IN and we must take two lines for space of a Key-word including a head line. When two IN correspond to a KW, excluding the head lines, we found that in one case out of ten the end numbers of both IN are the same, for example, 2301 and 3821; we take two lines in this one case and one line in other nine cases. We get 1.1 lines on an average.

In the same way, we get average lines $\bar{L}(n)$ for a Key-word corresponding to n INs as

$$\bar{L}(n) = \frac{1}{10^n} \sum_{L \geq n/10}^{n} \left(\sum_{a_1} \cdot \sum_{a_1} \cdots \sum_{a_i} \cdot {}_{10}\mathrm{P}_{a_1} \cdot {}_{a_1}\mathrm{P}_{a_2} \cdots {}_{a_{i-1}}\mathrm{P}_{i} \right) L$$

where

$$a_1 + a_2 + \cdots + a_i = n\,, \qquad 10 \geq a_1 \geq a_2 \geq \cdots \geq a_i$$

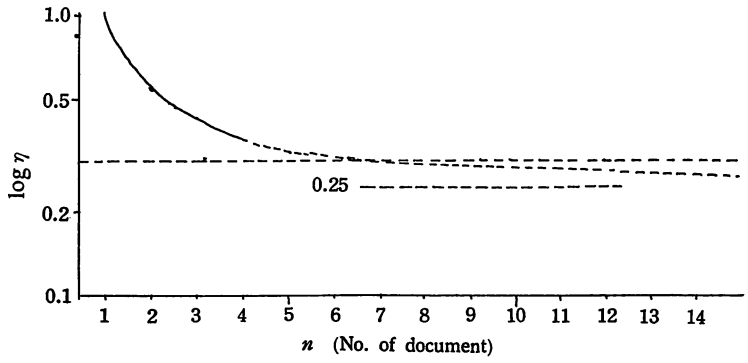

Fig. 2. Presumed value of η (average lines per document).

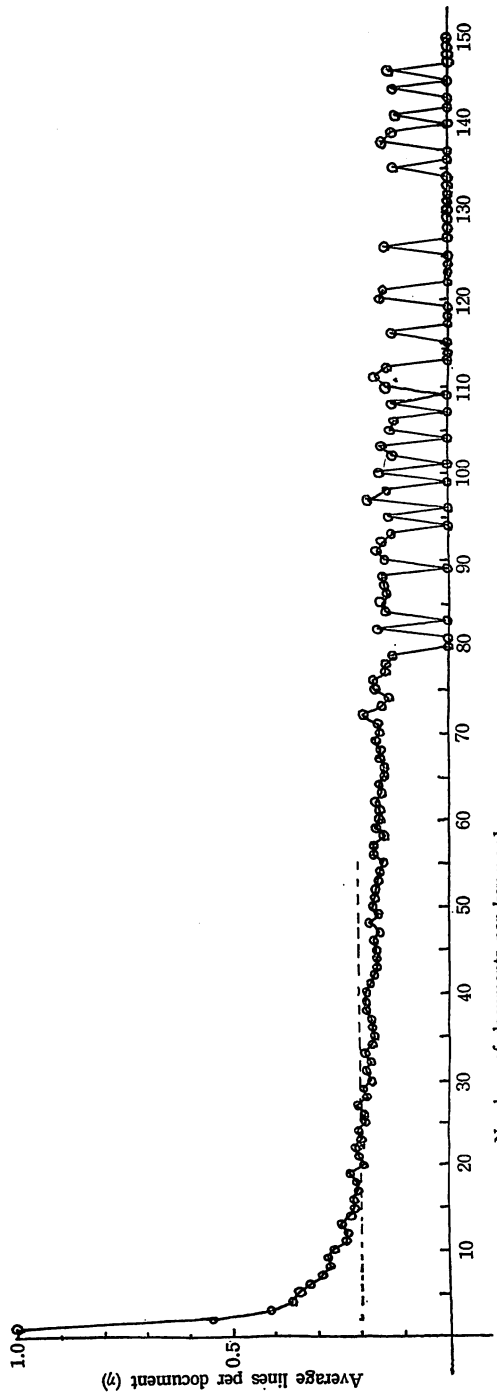

Fig. 3. Numbers of documents per Key-word and average lines per document.

Table 4. DN per Key-word and average lines per document.

n No. of documents per Key-word	η ($\bar{L}(n)/n$) (value measured)	η (value presumed initially)
1	1.00	1.00
2	0.55	0.55
3	0.41	0.43
4	0.36	0.40
5	0.35	0.36
6	0.32	0.33
7	0.29	0.30
8	0.27	0.30
9	0.28	0.30
10	0.26	⌐
11	0.24	|
12	0.23	|
13	0.25	|
14	0.23	|
15	0.22	|—0.28
16	0.22	|
17	0.21	|
18	0.21	|
19	0.23	|
20	0.20	⌐
21 to 30	average 0.20	|
31 to 40	average 0.18	|
41 to 50	average 0.17	|—0.25
51 to 60	average 0.16	|
100 to 110	average 0.14	|
150 to 160	average 0.13	⌐

Since this equation is difficult to solve, we obtained an indicator η of the average density of printed space on a graph by calculating $\bar{L}(n)$ for the case of $n = 1$ to 4, where

$$\eta = \frac{\bar{L}(n)}{n}$$

If $\eta = 1$, there is an IN in a line, *i.e.* only one-tenth of the space is printed. If $\eta = 0.1$, there are 10 INs in a line, and the space is completely filled.

If $n = 1$, we get $\eta = 1$, but if $n = 2$, $\eta = 0.55$, therefore it is better to delete the Key-word of $\eta = 1$ except for a few important descriptors. Thus the smaller value of η will result in a denser print, and the cost of the index will be reduced. We deleted Key-words of $\eta = 1$ to avoid blank spaces, except when those words were personal names or descriptors of a unique meaning.

Our estimation of η and its real value obtained by the index is shown in Fig. 2, 3 and 7 and Table 4.

XI. System Design

In compiling this index we decided very early to use a computer-oriented system of arranging the index file. In Dec. 1970, we chose a system known as "Input—Computer—Monotype" system.

According to this system, each document is first recorded in the computer memory, then the computor arranges each document in a file-order. This file is output not only on lineprinter in list form, but also on paper tape in punched form together with type picking code for monotype setting. These elements are described in detail below:

(1) Key-word coding and parity check

In general, it is very difficult to input a data such as Key-words of this index which are written in Chinese characters. To solve this difficulty, we developed a process by which a Key-word written in Chinese characters is converted to a number. For this, we developed a parity-check to guard against errors in analyzing, punching and handwriting. The parity check is achieved by writing the initial letter of the Key-word expressed in the Hepburnian system of romanization. This (1) provides direct correlation between the Key-word and KWN which indicates the full name of Key-word, (2) error input helps to make more easily detectable (26 characters of the English language as compared with 15 charactars in Japanese language and 0–9 numerical character). (3) When we check the alphabet of KWN with the initial letter of Key-word, the computor will check other errors in KWN (see Fig. 8). Regarding KW for Japanese names generally the initial letter of the personal name appears more frequently than that of the family name. So in the parity check we took the initial letter of the first name as written in Roman Alphabet. For chronological KW, the parity check code is the letter '*X*' in the case of a year, the letter '*Y*' in the case of year intervals.

(2) Compilation of dictionaries;

Two dictionaries used as a manual of computer input. One dictionary consists of KW terms arranged in the order of the Japanese syllabary; the other consists of KW terms arranged as element: in a facet or multi-dimensional classification. These KW terms are used to convert from KW to numerical bit when recording them in the computor memory.

(3) Flowchart (at initial stage)

First, we designed our "Input—Computer—Monotype" system

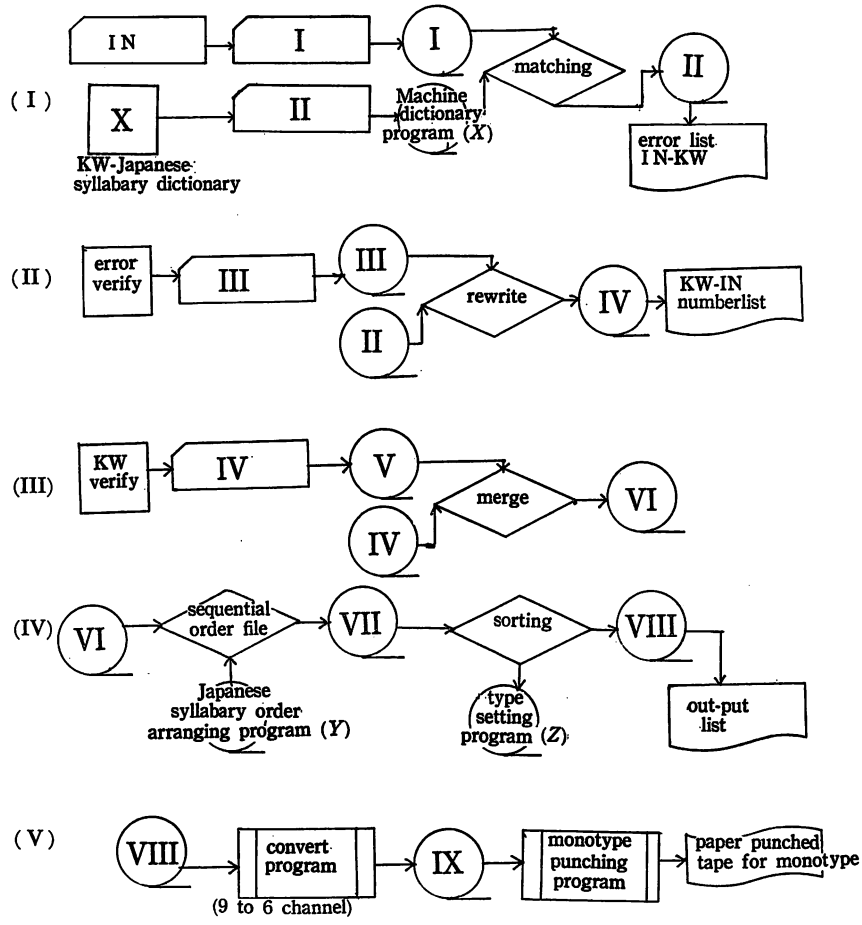

Fig. 4. Initial flow chart.

as follows: (Fig. 4)

(i) program card I → tape I and program card III → tape III will be the same. When an error occurs, error IN should not always be deleted unless it does not exist in the list.

(ii) the following steps are designed to add IN omitted from KW dictionary X on to a card

$$\text{CARD} \nearrow \begin{matrix} \text{TAPE V} \\ \text{TAPE IV} \end{matrix} \searrow \text{MERGE} \rightarrow \text{TAPE VI}$$

a: KW $(m, m', m'', \cdots)$, stored in the computor memory moves to or is duplicated on to KW (n), and KW (m) to KW $(n_1, \eta_2, \eta_3, \cdots)$

b: To delete KW $(m, m', m'', \cdots)$.

In the above case (a) (b), (α) ith, (α)i + 2th, $\cdots$ (α)i + jth merge to (α)ith, and (β) i move to k as merging. In both cases the same number is deleted through duplication. In addition, it is necessary to devise a program which can merge 1111 to 1152, for example, and a program which can merge (γ), (α) and (β). (Hereafter KW move is defined as MOVE, KW delete as DELT, and KW duplicate as DUPL)

(iii) Program (S) is needed when documents are to be retrieved from the computer memories, as shown bellow,

$$(DN_{KW_1} \cup DN_{KW_2} \cup DN_{KW_3}) \cap (DN_{KW_1'} \cup DN_{KW_2'} \cup DN_{KW_3'} \cdots) \cdots (S_1)$$

$$\left(\sum_{KW_i} DN\right) \cap \left(\sum_{KW_j} DN\right) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (S_2)$$

(4) Debug

There are two steps to produce a debug, that is:

$$KW_1\left(\xrightarrow[DUPL]{MOVE}\right) KW_2 \quad \left(\xrightarrow[DELT]{MOVE}\right) KW_3$$

First Step Second Step

(5) Out-put tape and format for type setting.

The main aim of this system is to ensure that out-put is directly converted to punched paper tape for use in monotype. In order to achieve this aim, out-put must be punched in 6 channel paper tape which codes two figures 00–99 depending on the type set box. As KWN in the index consist of four figures, from 1000 to 9999, the type setting code for our index requires a 20 code (10 + 10) to indicate KWN.

(6) Flow chart in operation.

Final results of the compilation of the index by computor are shown in Figure Fig. 5-1, 5-2. The program consists of five sub-programs named as SAKUIN 1–5, with 2,200 job-steps. Program language is COBOL, and the Computer used is TOSBAC 3400/41. Program documents are as follows, ([ ]$\cdots$ indicates a sub-program):

(a) Compiling data for a KW dictionary.

(b) Arranging KW dictionary according to the classification list with facets. [SAKUIN 1]

(c) Matching IN data (after IN modifier to up-date IN tape) with IN classified dictionary. [SAKUIN 2]

(d) Printing out an error list, which appears as a date-sheet written by manual, from the IN tape (Fig. 8).

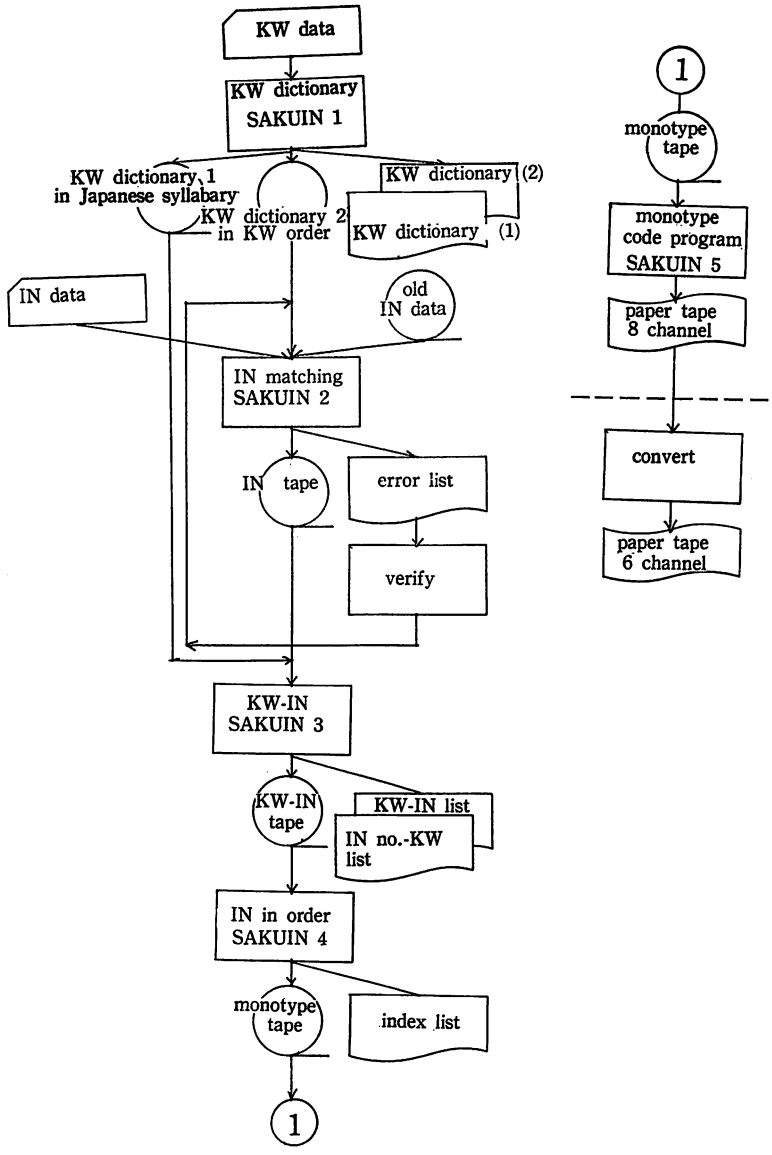

Fig. 5-1. Final flow chart.

(e) Using on error list to modify IN data and merge to IN file.

(f) Matching modified IN tape with IN Japanese syllabic dictionary, then outputting KW-IN list as an Index text. [SAKUIN 3] In this case, IN-KW lists are numbered for statistical use.

(g) Arranging INs, of the KW order-IN type list in numerical order [SAKUIN 4] whth this, the text index list is complete. Tape is

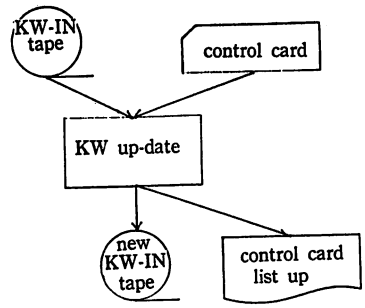

Fig. 5-2. Up-date file of KW-IN tape.

then used as a convert-tape for punching of monotype setting.

(h) By the code-convert sub-program [SAKUIN 5], the above tape is converted to digital code for monotype.

(i) Finally, the tape out-puts in punch 6 bit code for monotype from 8 bit tape (FACOM 230–50).

(j) In addition, a modifier (DUPLE, MOVE, DELET) is used as up-date KW-IN tape (shown in Fig. 5-2) and steps in the modification are inputed by a control card, to produce a new KW-IN tape. Control card-lists are printed out when a modification appears.

XII. Analyzing and Indexing of Historical Documents

(1) Analysers of documents

A large number of analysers were necessary to prepare an index as

Table 5. Consistency check on analysers.

Analyser	Total No. of KW selected from 4 documents	No. of correct KWs	
		Statistical method	Complementary method
A	40	20	25
B	35	17	21
C	33	22	17
D	32	18	19
E	41	16	15
F	33	20	18
G	36	21	19
H	27	17	16
I	45	22	20
J	30	17	17
K	21	16	15
L	25	16	18

Table 6. Errors of analysers, output by computor.

Volume	No. of errors per document			Total
	1	2	3	
1	3	3	3	9
2	5	—	—	5
3	12	—	—	12
4	12	—	—	12
5	17	1	—	19
6	6	—	—	6
7	9	—	—	9
8	12	—	—	12
9	1	1	—	3
10	8	—	—	8
11	16	1	—	18
12	6	2	—	10
13	12	—	—	12
14	24	1	1	29
15	12	—	—	12
16	—	—	—	0
17	3	—	—	3
18	28	3	—	34
19	11	—	—	11
20	18	3	—	24
21	16	—	—	16
22	28	2	—	32
23	12	—	—	12
24	17	3	—	23
25	18	—	1	21

comprehensive as ours. They had to be graduates of a college of technology, thus having a basic knowledge of science and technology and they should also be trained librarians, having some experience of indexing and classification. We decided to assign two volumes per analyser.

(2) Consistency check of analysers

Since analysers have different backgrounds and experience, the analysed results may differ widely. To guard against such differences and to measure personal error we devised a consistency check for all analysers.

(3) There are two methods of devising a consistency check. The first is a statistical method, which examines the KWs selected when the same document is distributed to all analysers. KWs selected by many

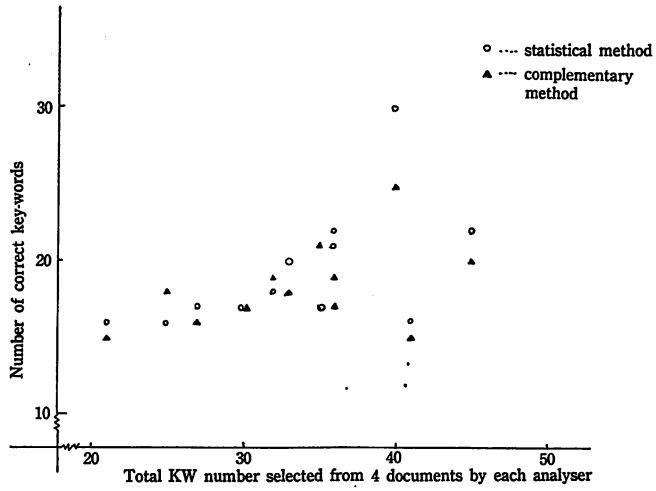

Fig. 6. Total correlation diagram of KW number and number of correct KW selected by each analyser from 4 documents.

analysers are more valuable than those selected by few or by no analysers. Those selected by few or no analysers are referred to as "Noize."

Arranging the selected KWs in the order to frequency of selection we set a boundary line at the KW where the summed frequency of the KW of the highest frequency reaches half of the total. By this method, the range of high frequency KWs is too broad. Therefore we need a more rigorous method of check consistency.

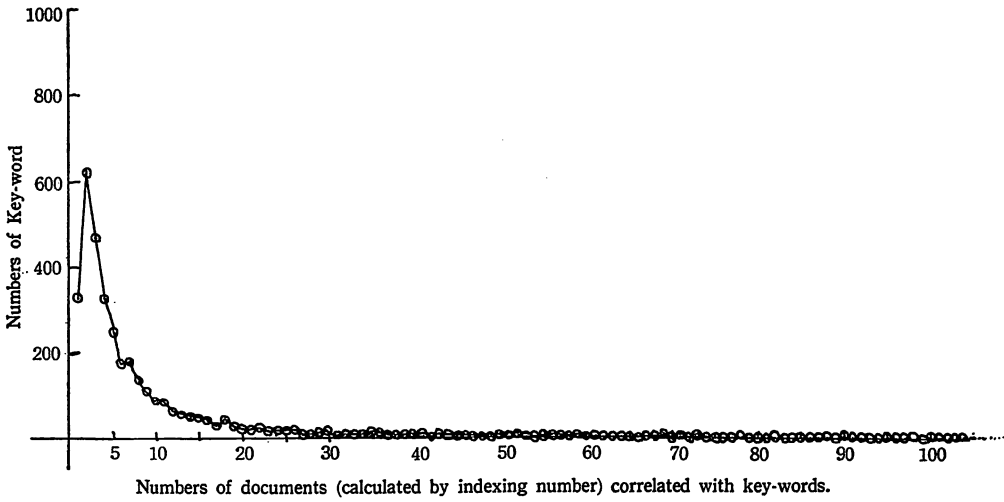

Fig. 7. Numbers of documents (calculated by indexing number) correlated with Key-words.

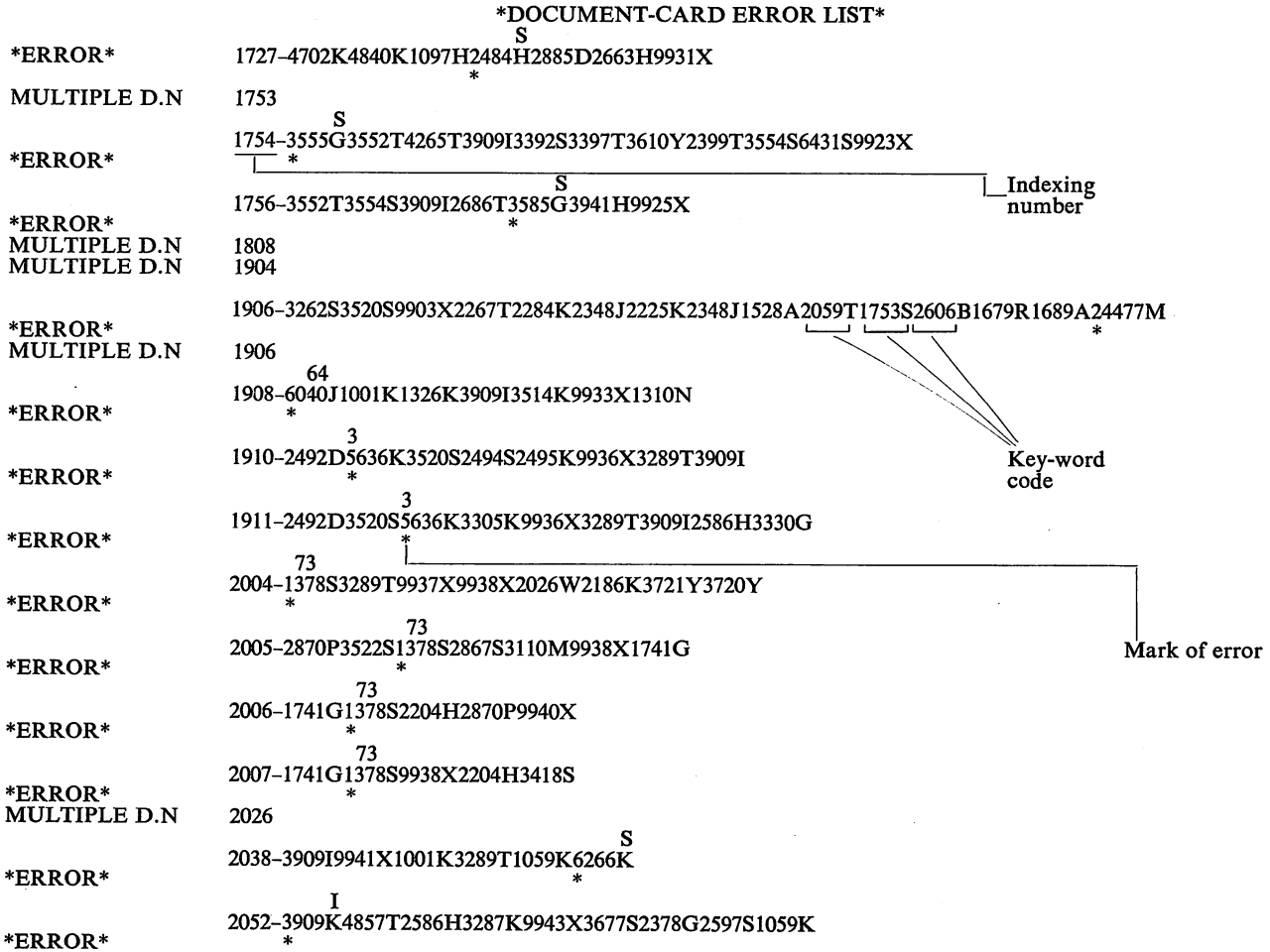

Fig. 8. Error list output by computor.

Regional number of classification		Key-word code
217	航路 (←水路)	2383K
217	倉庫	2384S
217	エレヴェーター	2385E
217	輸送機械	2386Y
217	交通機関	2387K
218	建築 (⇨504, 509, 609)	2388K
218	(建築論→建築)	2388K
218	(建築設計→設計)	2229S
218	施工	2389S
218	建築様式	2390K
218	洋風建築	2391Y
218	木造建築	2392M
218	石造建築	2393S
218	煉瓦造建築	2394R
218	鉄筋・鉄骨コンクリート建築	2395T
218	建築構造	2396K
218	柔構造	2397J
218	剛構造	2398G

Fig. 9-1. List of classified Key-words.

Regional number of classification		Key-word code
323	科学技術審議会	3483K
324	科学技術振興 (←科学技術援助)	3514K
324	科学技術政策	3513K
322	科学技術庁	3400K
530	(科学教育→科学技術教育)	
322	科学教育局	3445K
126	科学教育研究協議会	1315K
322	科学局	3444K
121	科学計測研究所	1157K
121	化学研究所	1170K
133	化学工学	1370K
263	化学工業	2875K
352	化学工業協会	3767K
323	化学工業調査会	3488K
352	化学工業調査会	
127	科学史	1327K
100	(科学者→研究者)	
403	科学者憲章	3930K
607	科学小説	4665K
323	科学審議会	3477K
126	科学振興調査会	1311K
603	科学精神	4632K
107	(科学動員→研究動員)	
126	科学動員協会	1312K

Fig. 9-2. List of Key-word in the order of the Japanese syllabary.

Table 7. Data from all 25 volumes of "Taikei".

Volume	Indexing number (IN)	No. of Documents	Number of KW in each volume																Total	No. of Key-words per document
			Science & technology	Industry & manufacture	Politics & economics	Social & labour problems	Medicine, education	Philosophy, thought, religion, & arts	Events	Geographical names	Japanese personal names	Foreign personal names	Chronology in general	Edo era	Meiji era	Taisho era	Before War II	After War II		
			1001K ~ 2171S	2201G ~ 3013M	3101S ~ 3787Y	3901S ~ 4125N	4201S ~ 4506S	4601M ~ 4714S	4801S ~ 4899S	5000N ~ 5201S	6001K ~ 6970S	7001A ~ 7167G	9601C ~ 9749S	9800X ~ 9867X	9868X ~ 9911X	9912X ~ 9925X	9926X ~ 9945X	9946X ~ 9970X		
1	1001~1257	257	504	309	367	260	73	35	10	82	285	24	96	27	242	0	0	0	2314	9.00
2	1301~1517	217	321	263	276	349	154	49	47	61	195	26	27	0	180	21	2	0	1971	9.08
3	1601~1813	213	335	437	338	240	52	33	42	43	125	9	44	0	7	163	58	0	1926	9.04
4	1901~2129	229	243	397	551	149	42	14	23	14	53	0	4	0	0	1	245	3	1739	7.57
5	2201~2418	218	351	347	676	266	47	18	16	7	44	1	18	0	0	0	17	226	2034	9.33
6	2501~2687	187	423	164	241	217	47	246	16	28	161	48	12	21	96	13	35	33	1801	9.63
7	2701~2852	152	490	99	203	113	96	34	11	235	31	71	213	23	47	27	68	35	1796	11.86
8	2901~3159	259	885	72	210	135	489	36	1	59	133	49	72	41	180	0	0	0	2362	9.12
9	3201~3405	205	335	35	89	138	530	16	3	17	129	18	37	0	127	76	10	0	1560	7.61
10	3501~3707	207	361	28	162	204	690	40	1	24	112	6	5	0	0	1	97	113	1844	8.91
11	3801~4029	229	572	121	116	76	99	33	166	228	164	9	32	12	74	35	60	112	1909	8.34
12	4101~4363	263	1223	107	59	205	57	39	24	37	456	32	91	5	36	15	98	127	2611	9.93
13	4401~4598	198	1105	94	36	87	14	16	13	27	276	63	15	1	52	34	103	59	1995	10.85
14	4601~4949	349	1941	85	178	174	37	6	50	213	265	41	43	17	151	40	87	108	3436	9.85
15	5001~5194	194	978	66	39	121	96	73	15	26	175	34	25	4	79	23	35	35	1824	9.40
16	5201~5373	173	129	573	359	84	210	11	29	193	110	25	64	1	63	19	30	43	1943	11.23
17	5401~5675	275	324	713	351	256	113	192	14	190	287	79	103	25	132	32	29	16	2856	10.38
18	5701~5908	208	318	740	218	105	19	14	10	59	137	21	51	11	45	32	96	66	1942	9.37
19	6001~6240	240	319	689	174	97	6	8	21	80	158	11	48	14	90	33	43	71	1862	7.79
20	6301~6511	211	360	530	127	106	16	22	22	85	123	6	72	9	74	28	40	34	1654	7.89
21	6601~6843	243	677	344	85	75	17	19	8	41	133	11	58	16	55	28	73	57	1697	6.98
22	6901~7131	231	624	916	306	166	55	1	25	170	140	34	46	13	237	12	5	2	2752	11.91
23	7201~7349	149	561	570	85	38	12	1	45	78	81	8	80	0	23	59	73	16	1730	11.61
24	7401~7553	153	283	47	178	154	362	14	19	87	140	28	4	22	182	4	0	0	1524	9.96
25	7601~7790	190	436	69	219	227	470	26	37	55	53	1	0	0	1	47	78	96	1815	9.55
Total		5450	14098	7815	5643	4042	3803	996	668	2139	3966	655	1260	262	2173	743	1382	1252	50897	Average 9.34

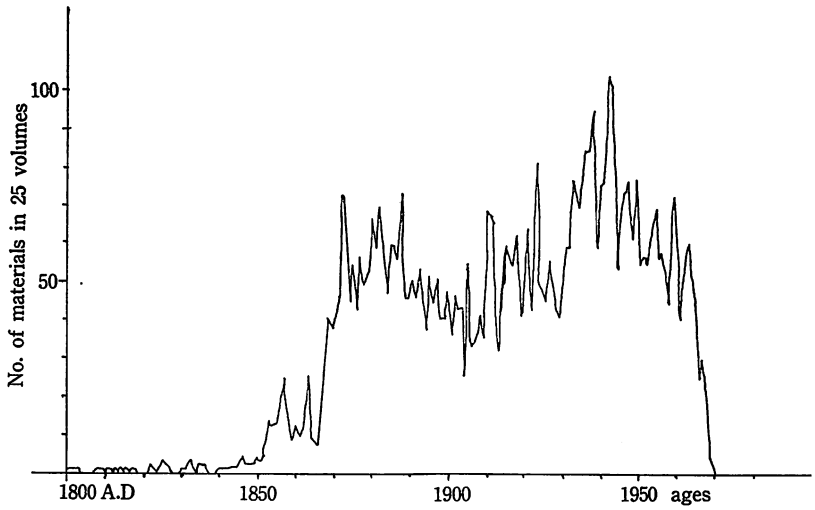

Fig. 10. Chronological distributions of materials in 25 volumes.

The second method, which is a complementary method, investigates how many KWs selected by an analyser match the KWs we have already selected as correct KWs for contents of the documents. Then we count matching KW selected by each analyser.

We recognized that both statistic and complementary, counts are not based on how many KW are selected by an analyser, but how many correct KWs are matched according to their choice. The result of the counts are shown in Table 5, 6.

There are some tendencies or characteristics of individual analysers which affect the number of KW selected and the number of KW matched thus affecting the total number of KW. This tendency is indicated in Fig. 6.

After the examination of the consistency check, we asked analysers to analyse each of the volumes.

XIII. Duration

The work of compiling, the thesaurus began on 9th October, 1968, and we spent 3–4 hours of voluntary work in one evening each week, with 2–4 members attending a meeting.

It took nine evenings, a total 24 persons' evenings, for the work of stage 1, and 56 persons' evenings for stage II.

In the Stage III, a member of a graduate course in librarian science listed words for most of the cards, and his work amounted to about 58 persons' evenings.

It took 9 evenings and one day and night, a total of 30 persons' evenings, for Stage IV; and 14 evenings and 3 days and nights, a total of 64 persons' evenings for Stage V.

Key-words	Indexing Numbers									Regional number of classification
				6463						*7038*
元老院										[322]
		5441					*5236*	*5237*		
小石川植物園										[538]
					1194	*5025*				
					5024					
小売店										[341]
				5593						
高圧送電 (⬅超高圧送電)										[227]
		6081	*1712*		*1714*			*6077*	*6078*	*6149*
			6072		*6204*			*6087*	*6088*	
耕耘										[241]
			6942		*7064*	*6945*	*6906*	*6997*		
			7212				*6996*			
耕耘機・トラクター										[256]
				7213	*6924*	*7205*			*1678*	
									7218	
公園 (⇨171)										[513]
	5230	*5241*	*5232*	*1753*	*5244*	*5235*				
	5290	*5331*	*5332*	*5243*	*5294*					
	5300				*5344*					
	5330									
航海										[217]
		1011	*4862*			*1335*	*4616*	*2907*		*2919*
						2905	*4656*	*4837*		*4609*
郊外										[505]
	5260		*5302*		*5344*		*5256*		*4778*	*5259*
	5300								*5258*	
公害										[706]
	1370	*1371*	*1372*	*1723*	*1724*	*1725*	*1726*	*1727*	*1368*	*1369*
	1450	*1451*	*1452*		*4014*	*6205*	*7776*	*7777*	*1728*	*1449*
	6840	*6841*	*5372*		*4024*	*7775*				*4019*
公害病										[518]
		7531				*7775*		*7777*		

Fig. 11. Sample of Index.

From the Stage I to the Stage V the total was 232 persons' evenings, excluding simple tasks such as rearranging words in the order of the Japanese syllabary, making fair copies, and making duplicates. The classified Key-word list was completed on 23rd January, 1971, and the Key-word list in order of Japanese syllabary was made by re-arranging terms of the classified list. We had several meetings to explain and amend the Key-word list to analysers from 6th March, 1971.

When the analysers' work was completed, the data for indexing documents were handed over to JUSE computor centre. The computor of TOSBAC 3400–41 ran for 10⅔ hours and a monotype machine of DAI-ICHI-HOKI SHUPPAN Co., Ltd. set up types for a week, nearly 60 hours.

The number of pages of this index was 358, almost equal to the pages estimated statistically in Fig. 2, 3 at the beginning of the computer in-put stage (the difference is about 8%).

XIV. Conclusion

We show the frequencies of Key-words in each volume. It does not indicate the distribution of analysers so much as the difference of documents contained in each volume (Table 7).

We express our thanks to Dr. NAKAMURA Hatsuo, professor of Keio-Gijuku University; Mr. YAJIMA Keiji, of JUSE company, who assisted in programming and running the electronic computor; Mr. ISHIKAWA Toru, DAI-ICHI-HOKI SHUPPAN Co. Ltd. Publisher of "*Taikei*;" and the members who, under difficulties, analysed the documents of the "*Taikei*."

Résumé

THESAURUS IN HISTORY OF SCIENCE

—Editing an index for Nihon Kagaku-gijutsu-shi Taikei
(History of Science and Technology of Japan)
using a computer—

A thesaurus of *Nihon Kagaku-gijutsu-shi Taikei* has been compiled for indexing historical documents of science and technology of Japan. Various features of the thesaurus were considered, and classification system involving facets of personal names, events and chronological data was employed.

All documents of *Taikei* were analysed, and the data arranged according to the requirements of the thesaurus.

Taikei includes 2,720 general Key-words and 1,064 personal names. The computer program consisted of 2,200 steps in Cobol language and the computer ran 10 hours and 40 minutes. A direct output was obtained from monotype composition code input. It took two and a half years to complete the work.