# APTS Code: an Identity Code Applicable to Various Types of Documents

Isao Asai

College of Engineering, University of Osaka Prefecture
Gakuen 1-1, Sakai, Osaka, 593, Japan

Abstact

Pinpoint search, directly to search only a specific document, got necessary. In this case, we use the accession number of database. However, it differs from every database, and it is difficult to know its number. If all databases adopt a common identy code against a document, pinpoint search may get possible. This paper aims to propose a new method to get a unified code applicable to various types of documents: journal articles, conference articles, books, and so on. It is desirable to get a code with an easy method for anyone, anywhere, and any time. That is, what you see is what you code. A new code, named with APTS code from the capital, composes it from the author, publication year, title and source. I discuss it about a rule for getting APTS code and the application fields.

Keywords: pinpoint search, accession number, database, coding, identity code, APTS code.

## 1. Introduction

The package in the distribution of documents is moving from journals to an individual article. The journal that collects articles on a specific topic performs a large part for the distribution of articles today also. However, each researcher requests an article by online information services and inter-disciplinary research. In such a case, pinpoint search that takes out only one article is necessary.

Various databases, such as current contents, abstracts, and fulltexts, load to a content of articles simultaneously with the publication of journal. For examples, Information Processing & Management has a list about twenty four databases recorded by a back cover.

Each database gives a unique accession number to order of reception. If twenty four kinds of databases deal with the same article respectfully, the different twenty four kinds of identity code may be assigned. Moreover, each database has a unique attribute, content, and expression against the same article. Consequently, it is impossible to identity as they are the same document, about a document that scatters to many databases.

In the field of information, an international standard code exists. That is ISSN for journals, and ISBN for books. Though these publisher oriented code contribute to be large at the distribution of journals and books, researchers don't have an interest enough. As they are a code to pass for only journals and books, such a code that can use commonly for various types of documents is necessary.

For examples, the database of ISI's company deals with references instead of abstracts. References include various types of documents such as journal articles, conference articles, books and reports. The next shows ISI's code of references.

SMITH LC (LIBRARY TRENDS, V30, P83, 1981).

The code of a document composes it from a bibliographic item about author and source. Then, one document expresses it with 30 - 50 bytes. Citation index to obtain by transformation of references give us very useful information. Because there is a few case that raw data uses omission type of journal name, it needs an attention to use the code[1].

In this paper I propose a new method to get a unified code applicable to various types of documents: journal articles, conference articles, books, and so on. It is desirable to be able to get a code with an easy method for anyone, anywhere, and any time. That is, what you

see is what you code. Then I examined a method that makes a code from fundamental bibliographic items of document. A method that I state here uses to make an author's article[2] greater in quality.

## 2. Development of APTS code for identifying documents

### 2.1 The configuration of APTS code

APTS code that I propose newly composes it with the fundamental bibliographic items of document because the acquisition of data is easy. By anyone, anywhere, we can make the code of document. We should pay attention to that references compose it with fundamental bibliographic data.

The number of documents published by one year is 4 - 5,000,000 and it increases yearly. A code against 100,000,000 documents is eight columns necessity with a figure only. Moreover, it is four columns necessity of the western calendar. Consequently, it is total twenteen columns necessity. However, it takes cost to attach a number for all documents, and timelag occurs. Therefore, this plan is unrealizable.

Here, I considered an identity code that composes it from author, publication year, title, and source, with each four columns. Therefore, it is with total sixteen columns. And it named with APTS code from the capital.

Figure 1 shows the configuration of APTS code with sixteen columns. And Figure 2 shows APTS code on references[1-3]. The following sections explain it every bibliographic item about a rule for making APTS code.

| Author | | | | Publication Year | | | | Title | | | | Source | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| **A** | **S** | **I** | **I** | **1** | **9** | **9** | **4** | **A** | **C** | **I** | **C** | **E** | | | |

Fig. 1 APTS code with sixteen columns

| | Author | Publication Year | Title | Source |
|---|---|---|---|---|
| Ref.1 | **PENO** | **1994** | **IBRF** | **5031** |
| Ref.2 | **ASII** | **1993** | **DLRD** | **E273** |
| Ref.3 | **ASII** | **1987** | **RDDI** | **E001** |

Fig.2 APTS code on references[1-3]

### 2.2 Author's coding

I made the data file with 2,570 author's names to examine author's code. An average number of character in family name was 6.5. I used JGAWK of text processing language for data analysis. I made a rule for getting code, made the program to correspond to it, and obtained it with the duplicated number of code and the total number of authors. The next shows the examined rules and the results.

(1) First four characters of family name.

Duplications was 401 and the total number of authors was 1,162.

(2) First three characters of family name and first character of first name.

Duplications was 255 and the total number of authors was 580.

(3) First two characters and last character of family name, and first character of first name.

Duplications was 238 and the total number of authors was 515.

I adopt it, as considering that as author's rule, as the third method is the most superb. Figure 3 shows the coding rule of author. We express author's name with full spell of family name, comma, capital of first name and period. In this case, we apply rule of the first two characters of family name and two characters to clip a comma, in principle.

By the name of author, an exceptional treatment is necessary. Figure 4 shows the examples of normal and exceptional author's coding.

| Column | Coding rule of author |
|---|---|
| 1 col. | first character of family name |
| 2 col. | second character of family name |
| 3 col. | last character of family name |
| 4 col. | first charcter of first name |

Fig. 3 Author's coding rule

| Author's name | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **Bradford, S. C.** | **B** | **R** | **D** | **S** |
| **Henry Small** | **S** | **M** | **L** | **H** |
| **Yu, C. T.** | **Y** | **U** | **U** | **C** |
| **MacRae, D.** | **M** | **A** | **E** | **D** |
| **Sparck Jones, K.** | **S** | **P** | **S** | **K** |

Fig. 4 Representative examples of author's coding

### 2.3 Publication year's coding

A code of publication year uses four columns of the western calendar. Publication year of a position is just behind an author, or is for a portion behind the source. Figure 5 shows publication year's coding rule.

If publication year's code expresses with lower two

columns of the western calendar, I can shorten two columns of code. However, a miss is a few compared with four columns. Moreover, the code of four columns is available for twenty one century. Figure 6 shows the examples of publication year's coding.

| Column | Coding rule of publication year |
|---|---|
| 5 - 8 col. | western calendar |

Fig. 5 Publication year's coding rule

| Publication Year | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| Bradford, S. C. (**1934**). | **1** | **9** | **3** | **4** |
| ..,44-1,221-238 (**1993**). | **1** | **9** | **9** | **3** |

Fig. 6 Examples of publication year's coding

## 2.4 Title's coding

Similarly to the case of author, I made the data file with 2,600 titles subjects to examine coding. Then, I made various rules and examined it. An average number of character in title was 70.1, and an average number of words was 9.7. In case of excluding a word inside three character, an average number of word was 6.6.

I kept and converted a special character inside a title to a blank. Consequently, a word with a slash saw and performed it with two words.

Title's coding is to choose four characters from average 70 characters. We can think about many methods for its selection. I requested a number of duplicated code and a total number of titles at a following rule.

(1) Capital of four words from the beginning.

Duplications was 220, and the total number of titles was 551.

(2) Capital of four words from the beginning at the title excluded unnecessary words.

Duplications was 108, and the total number of titles was 259. Here, unnecessary word defines it with word inside two characters and five words: and, for, its, the, and with.

(3) Capital of four words from the beginning at the title excluded words inside three characters.

Duplications was 80, and the total number of titles was 198.

I adopted it as considering third method as a rule. Figure 7 shows title's coding rule. In the case of from one to three words, an exceptional treatment is necessary for coding title. For these cases, I decided to make up several insufficient character from last portion at last word. This rule is a nature method from a movement of an eye to data. Figure 8 shows the examples of title's coding.

| Column | Coding rule of title |
|---|---|
| processing in advance | convert special character inside a title to a blank |
| def. of word | word more than four characters |
| 9 col. | first character of first word |
| 10 col. | first character of second word |
| 11 col. | first character of third word |
| 12 col. | first character of fourth word |
| one to three words | make up insufficient character from last portion at last word |

Fig. 7 Title's coding rule

| Title | 9 | 10 | 11 | 12 |
|---|---|---|---|---|
| The **S**tructure of **S**cientific **L**iteratures. I: **I**dentifying and Graphing Specialties. | **S** | **S** | **L** | **I** |
| **L**inguistics and **I**nformation Science. | **L** | **I** | **S** | **E** |
| **C**itation **A**nalysis. | **C** | **A** | **I** | **S** |
| **B**ibliometrics. | **B** | **I** | **C** | **S** |

Fig. 8 Representative Examples of title's coding

## 2.5 Source's coding

It is getting the code of source data hardly to treat it first. I examined to the description of when many documents reference one document. The description of author, publication year, and title is generally similarly. However, the description of source is delicately different. Particularly, as we use an omission type for the source, an specialist only is comprehensible.

The description of source differs from data every time. I made approximately 5,300 source data and investigated a data other distribution. For the result, periodicals for 59 %, conferences and editorials for 17 %, books for 19 %, and the remainder for 5 % was. Consequently, I do getting the code of source with three kinds of methods.

### 2.5.1 Periodicals

It is necessary for getting a code to use a data that any document include in. Almost all documents have journal name, volume, and a starting page. Because there is the case of journal name with an omission type,

it does not contain code with journal name. The code of periodicals makes it with lower one column of volume, and lower three columns of starting page. Figure 9 and 10 shows coding rule and examples of periodicals.

| Column | Coding rule of periodicals |
|---|---|
| 13 col. | lower one column of volume |
| 14 - 16 col. | lower three columns of starting page |

Fig. 9 Coding rule of periodicals

| Periodicals | 9 | 10 | 11 | 12 |
|---|---|---|---|---|
| Journal of Documentation, **28**, **11**-21 (1972) | **8** | **0** | **1** | **1** |
| Nature, 22**1**, 1**205**-1207 (1969) | **1** | **2** | **0** | **5** |

Fig. 10 Coding examples of periodicals

### 2.5.2 Conferences and Editorials

The abstracts and articles of proceeding, and books editorial articles hit here. Behind a title, it lists In...( Eds.), In Proceeding, and so on, and then continue the place of holding, page, and publishing company. This case also disregards the name of proceeding and books, and expresses a code using character "E" and the beginning three columns of starting page. Figure 11 and 12 shows coding rule and examples of conferences and editorials.

| Column | Coding rule of editorials |
|---|---|
| 13 col. | character "E" for editorials |
| 14 - 16 col. | lower three columns of starting page |

Fig. 11 Coding rule of conferences and editorials

| Conferences and editorials | 9 | 10 | 11 | 12 |
|---|---|---|---|---|
| Proceedings of 43rd FID Conference, (Sept. 1988), **47**-54. | **E** | **0** | **4** | **7** |
| In M. Dillon (Ed.), .., (pp.**159**-168). (1991). NY: Greenwood Press | **E** | **1** | **5** | **9** |

Fig.12 Coding examples of conferences and editorials

### 2.5.3 Books and the others

The bibliographic data of books describes it with author, publication year, title(italic), publication place: publication company. The code of books composes it with character "B" and the first three characters of a principal word in the publishing company.

The first column of code regarding to report and dissertation, similarly to the case of books, indicates the type of data. It uses character "R" for report, "D" for dissertation, and "Z" for the others. Then three columns of remaining uses the first three characters of a principal word in the publishing company. Figure 13 and 14 shows coding rule and examples of books and the others.

| Column | Coding rule of books and the others |
|---|---|
| 13 col. | character "**B**" for books,<br>character "**R**" for reports,<br>character "**D**" for dissertation,<br>character "**Z**" for the others |
| 14 - 16 col. | first three characters of a principal word in the publishing company |

Fig. 13 Coding rule of books and the others

| Books and the others | 9 | 10 | 11 | 12 |
|---|---|---|---|---|
| Amsterdam: **Elsevier** | **B** | **E** | **L** | **S** |
| Final report. **Case** Western Reserve University | **R** | **C** | **A** | **S** |
| Doctoral dissertation, University of **Cal**ifornia at Berkeley | **D** | **C** | **A** | **L** |

Fig. 14 Coding examples of books and the others

## 3. Discussions

We examine it about the quantity of documents available using APTS code. The configuration of author is alphabet. If considering that it is every column by ten kinds, the number of combinations becomes to 10,000 kinds, because author's code is four columns. Again, similarly to author, the number of combinations in title's code is 10,000 kinds. Source's code composes it with numerical value of three columns and if doing, the number of combinations is 1,000 kinds. Consequently, APTS code can be enough useful, even though researchers in the world producted yearly a hundred billion documents.

For the next, I enumerate an objective criterion of evaluation to obtain coding rule.

1) Make for easy rule, as it is possible, to prevent miss.
2) Make the number of columns of code to a few.

3) Correspond to the quantity of document for several 100 years.
4) Make code automatically from existing databases.
5) Open for many databases.
6) Don't make duplicated code.

Here, to examine coding method, I made the code of 5,300 documents from references of a periodical for four years. Then, I requested documents that four kinds of subcode all fits to. Again, I requested documents that one kind of subcode differs from. I made a program and examined the place where an error occurs. Then, variously I repeated trial and error and obtained above rule.

The code of one hundred documents per one hour was able to make it from a bibliographic data. In case that a bibliographic data itself is wrong, we can't make a correct code. We must be avoid to have an error for getting code.

An error can easily occur, in case that it is exceptional. I state the case that it occurs.
a) Family name of author is complicated.
b) We omit subtitle and long title.
c) We mislead a code for the place of publication and university name.

Because I didn't employ journal name and proceeding name, coding was able to make it rapidly moreover, easily. Again, because the total number of code is a few, I was able to manage it with personal computer. Generally speaking, I obtained very satisfactory results.

## 4. Application fields of APTS code

As articles has identity code, a new world appears up to this time. In this section, I consider it about an application fields of APTS code.

### (1) Making of document directory.

At the design of database, a tendency that the attribute of document increases sees. Because the total number of byte per document increase, the operation of database deteriorates. Opposite, for the attribute of document composed of author name, publication year, title, and source, database becomes compact, and the operation can easily do.

Many current database treat documents only since 1960's. However, database containing documents published to old generation is necessary. It is realizable if being the case, database to have an only fundamental bibliographic item of document. For a unified document code if being given, the value of utilization gets expensive.

### (2) Adoption of all databases.

If fulltext database advances, a search to a specific document gets necessary. If all databases have unified document code, pinpoint search will be possible, and document delivery will be easy. APTS code is available by anyone, as being able to make it from a bibliographic data of document. The value of database to have unified document code is expensive.

### (3) Application to references.

The quantity of documents increases from year to year. Then the number of document per references increases. To settle a circulation problem of documents, though restricting the character number of paper also, thereafter, the number of document per references may become a question. As a measure to settle that, there is the utilization of APTS code.

### (4) Application to referation database.

Citations obtain by transferring references. Referation composes it from references, document itself, and citations[3]. Referation can form a new information search and knowledge base, as expressing the link between documents. As we employ APTS code, a referation database can make it easily.

## 5. Summary

Online information search is receiving conversion term with progress of CD-ROM and appearance of fulltext database. Then, information retrieval is shifting from scoop search to pinpoint search. The case, the part of identity code is large. APTS code as developed here is one powerful method. Again, the synergistic effect is large, if a current database has APTS code.

## References

[1] Persson, O. (1994). The intellectual base and research fronts of JASIS 1986-1990. *Journal of the American Society for Information Science,*45-1, 31-38.
[2] Asai, I.(1993). Development of large referation database: Introduction of APTS code for identification of scattered documents (pp.273-278). *Proc.29th JICST Annual Meeting, Tokyo, Japan: JICST (In Japanese).*
[3] Asai, I. (1987). "Referation" database for document information analysis (Vol.24, pp.1-5). *Proc.50th ASIS Annual Meeting, Boston: ASIS.*

International Federation for
Information and Documentation

# 47th FID

# Conference and Congress

## Finding New Values
## and
## Uses of Information

Sonic City Omiya, Saitama, Japan

October 5-8, 1994