Bibliographic File Integration System Using PYAT Identifiers

Matsui, Sachiko
University of Library and Information Science

Matsuda, Yoshiro
Institute of Economic Research, Hitotsubashi University

Asano, Chooichiro
Research Institute of Fundamental Information Science, Kyushu University

https://doi.org/10.15017/17172

出版情報：九州大学大学院総合理工学報告. 11 (4), pp.413-421, 1990-03-01. 九州大学大学院総合理工学研究科
バージョン：
権利関係：

九州大学総合理工学研究科報告
第11巻第4号413-421頁 平成2年3月

Engineering Sciences Reports, Kyushu University
(KYUSHU DAIGAKU SOGORIKOGAKU KENKYUKA HOKOKU)
Vol. 11, No. 4, pp. 413 - 421, March 1990

Bibliographic File Integration System Using PYAT Identifiers

Sachiko MATSUI*, Yoshiro MATSUDA**
and Chooichiro ASANO***

(Received November 30, 1989)

Identification of duplicate items in different bibliographic databases is one of the most important operations to construct large-scale databases for practical use. The PYAT identifiers are general terms for the techniques we developed for effective identification. P, Y, A, T are the components of the identifiers, such as publishing place, year, author name and title. The identification experiment clarified that components of identifiers give different effects on the identification efficiency: the success ratio in identifying duplicates by computer. Therefore, the identification efficiency thus acquired can be evaluated as useful data for composing a more effective identifier.

1. Introduction

Books and journals are easily identified if they are assigned unique identification numbers such as ISBN (International Standard Book Number) and ISSN (International Standard Serial Number). However, the literature for retrospective searches was mostly published before the introduction of these identification numbers.[1] Thus, a method of identifying duplicate items based on bibliographic descriptions must be developed. Such an identification method is necessary for compiling a database by integrating two or more existing machine-readable files.[2] This paper summarizes and evaluates the techniques which we have developed to identify duplicate items using identifiers based on bibliographic descriptions to suit the needs of the literature on social sciences.[3]

2. Design of PYAT Identifier and Definition of Identification Effectiveness

The basic idea of computer identification in this study is to compare character strings constituting identifiers automatically extracted from bibliographic descriptions of a bibliographic item (i.e., a book). We used author names, titles, and places and years of publication as the components of identifiers. We then evaluated the efficiency of these identifiers, i.e., how accurately they can discriminate duplicate items contained in two separate machine-readable files.

The data files used in the experiment are as follows:

1) The ECON (Economics) file, one of the subject files in the *Widener Library Shelflists*[4] database of Harvard University. It contains abundant catalog information on past

*University of Library and Information Science
**Institute of Economic Research, Hitotsubashi University
***Research Institute of Fundamental Information Science, Kyushu University

literature.

2) A machine-readable file compiled by the RINET Society, our research group, containing bibliographic items reviewed in the "Book Reviews" section of *Econometrica* (Journal of the Econometric Society: an international society for the advancement of economic theory in its relation to statistics and mathematics) issued in the United Kingdom since 1933). The file is referred to as the "EBR" file in this paper.[5]

Table 1 shows the outline of the two files. The ECON file contains 66,229 records covering the economic literature in general. The oldest date of publication goes back to 1581. On the other hand, the EBR file contains 864 records of academic literature closely related to econometrics and published from 1947 to 1969. Thus, from the ECON file, we have extracted records on the literature published in the same period as the EBR file. The number of records extracted is 23,202. We have not used subject criteria represented by classification entries in the ECON file, because econometrics is a method and not a subject.

The form of bibliographic descriptions has the most important role in the identification experiment. In the ECON file, the descriptions follow the local and compact cataloging rules for the *Widener Library Shelflists*. However, the description forms of the EBR file are similar to those of citations by researchers in their writings.

For this identification experiment, we have used four types of data elements contained in both the ECON and EBR files. They are **P**lace and **Y**ear of publication, **A**uthor name

Table 1 Outline of ECON and EBR files used in PYAT experiment

	ECON	EBR
printed version	Harvard University Library *Widener Library Shelflists Nos. 23—24: Economics and Economic Periodicals* (Harvard University Press, 1970).	Yoshiro, Matsuda & Sachiko Matsui, *Bibliography: Works on Econometrics—Reviewed in Econometrica 1951—1971* (RINET Society, 1973).
scope of subject	Economics in general.	Econometrics.
number of books	66,229.	864.
period of publication	1581—1973 (latest input to the database is April 8, 1974).	1947 — 1969, materials reviewed in *Econometrica* vols. 19—39 (1951—1971).
subfile for experiment	Extraction of 23,202 books (35% of total) published from 1945 to 1970 including materials without publication date. Languages are English (51%), German (15%), French (11%), Russian (7%) and others (16%).	All the 864 books in EBR database. Languages are English (77%), German (9%), French (8%), Russian (3%), Italian (2%) and others (1%).
description rules	Abridged local description rules. Author: main entry only (one person or institution). Title: short title (subtitle and series title are omitted). Imprint: place and year of publication only (publisher is omitted). Collation: omitted.	Description rules based on those in "book review" of *Econometrica*. Author: up to four authors. Title: including subtitle and series title. Imprint and collation: year of publication, pages, place of publication, and publisher.

and **T**itle. Thus, we call the identifier **PYAT**, and the identification experiment is referred to here as the **PYAT** experiment.

First, we define duplicates and identifiers for identification by computer. Next, we describe the components and the composition of **PYAT** identifier. Then, we introduce the definition of the "duplicates" in the **PYAT** experiment and two measures for the evaluation of identifiers.

1) Definition of duplicates and identifiers for computer identification

Suppose a bibliographic item x_i, belonging to a set $\boldsymbol{X}(x_i \in \boldsymbol{X})$, is described by a data element $z^{(i)}$, $i=1, \cdots, k$, and two subsets $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are selected from $\boldsymbol{X}$. Among $z_{1m}{}^{(i)}$ for x_{1m} belonging to $\boldsymbol{X}_1$ $(x_{1m} \in \boldsymbol{X}_1)$ and $z_{2n}{}^{(i)}$ for x_{2n} belonging to $\boldsymbol{X}_2$ $(x_{2n} \in \boldsymbol{X}_2)$, two or more bibliographic items, whose data elements contained in the definition of the duplicates (see 4) below) are identical, are regarded here as duplicates. This means that for $x_{1m}{}^*$ among x_{1m} and for $x_{2n}{}^*$ among x_{2n}, when $z_{1m}{}^{(i)*}=z_{2n}{}^{(i)*}$, $i=1, \cdots, l$, then $x_{1m}{}^*=x_{2n}{}^*$. Therefore, $x_{1m}{}^* \in \boldsymbol{X}_1{}^* \subseteq \boldsymbol{X}_1$, and $x_{2n}{}^* \in \boldsymbol{X}_2{}^* \subseteq \boldsymbol{X}_2$.

In the identification by computer, it is generally impossible to use all the data elements, $z^{(1)}, \cdots, z^{(k)}$, describing one bibliographic item, because the types of data elements contained in existing databases are limited and different from each other. Moreover, since their description rules are often not unified among databases, it is not effective to use the entire character strings constituting each data element for identification. For these reasons, first the types of data elements common to two subsets $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are selected, and then character strings extracted from a data element are assembled into an identification key. This key is defined here as an identifier $\hat{z}^{(i)}$. Since identification by such identifiers does not use all the data elements contained in identical items, the duplicates indicated by such identifiers are not necessarily the "real" duplicates.

In the experiment, one-to-one comparison of the identifiers was made between $\hat{z}_{1m}{}^{(i)}$ for x_{1m}, all the elements of $\boldsymbol{X}_1$, and $\hat{z}_{2n}{}^{(i)}$ for x_{2n}, all the elements of $\boldsymbol{X}_2$. Here x_{1m} and x_{2n}, satisfying $\hat{z}_{1m}{}^{(i)}=\hat{z}_{2n}{}^{(i)}$, are regarded as $\hat{x}_{1m}$ and $\hat{x}_{2n}$ $(\hat{x}_{1m} \in \hat{\boldsymbol{X}}_1 \subseteq \boldsymbol{X}_1, \hat{x}_{2n} \in \hat{\boldsymbol{X}}_2 \subseteq \boldsymbol{X}_2)$. If $\hat{x}_{1m}=\hat{x}_{2n}$, they are regarded as duplicates. However, there is no guarantee that $\hat{x}_{1m} \in \boldsymbol{X}_1{}^*$ and $\hat{x}_{2n} \in \boldsymbol{X}_2{}^*$. Therefore, $\hat{x}_{1m}$ satisfying $\hat{x}_{1m} \in \boldsymbol{X}_1{}^*$, and $\hat{x}_{2n}$ satisfying $\hat{x}_{2n} \in X_2{}^*$, are defined as "complete matches." They belong to a subset $\boldsymbol{M}$. Conversely, $\hat{x}_{1m}$ satisfying $\hat{x}_{1m} \notin \boldsymbol{X}_1{}^*$, and $\hat{x}_{2n}$ satisfying $\hat{x}_{2n} \notin \boldsymbol{X}_2{}^*$ are defined as "noise." They belong to a subset **N**. In addition, x_{1m} satisfying $x_{1m} \in \boldsymbol{X}_1{}^*$ and $x_{1m} \notin \hat{\boldsymbol{X}}_1$, and x_{2n} satisfying $x_{2n} \in \boldsymbol{X}_2{}^*$ and $x_{2n} \notin \hat{\boldsymbol{X}}_2$, are defined as "leakage." They belong to a subset **L**. These subsets, **M**, **N**, and **L**, contain different elements, depending upon the design of the identifier $\hat{z}^{(i)}$. The identification experiment aims to measure the effectiveness of various identifiers $\hat{z}^{(i)}$ consisting of various data elements with different lengths of character strings.

2) Components of the **PYAT** identifier (See **Table 2**)

P : first one character of the publication place.

Y : publication year, four-digit number.

A : first three or five characters of the last name of the author.

Table 2 Components and composition of PYAT identifier

	W34	TUM	TAO	PYA
type of identifier	Three or four title words	Title unmodified	Title, first article omitted	PYA (see below)
component P Y A	P: first one character of publication place. Y: publication year (four digits). A: first three characters of the last name of the author.			P & Y are as same as those of W34, and A is first five characters.
component T and numbers of words or characters extracted	From the beginning of each word (excluding article). Combinations are as follows: a: 3,1,1 e: 3,1,1,1 b: 3,1,2 f: 3,1,2,1 c: 3,2,1 g: 3,2,1,2 d: 3,2,2 h: 3,2,2,1	From the beginning of the first character, various length of strings are extracted as follows: a: 15 d: 45 b: 25 e: 55 c: 35	As same as that of TUM. First article omitted. a: 15 d: 45 b: 25 e: 55 c: 35	None.
combination of components	Eight kinds of combinations: PYAT, PAT, PYT, PT, YAT, AT, YT and T.			PYA.
number of identifiers	64 kinds (8 × 8).	40 kinds (8 × 5).	40 kinds (8 × 5).	One kind.
length of identifiers	13~16 characters.	23~63 characters.	23~63 characters.	10 characters.

T: (1) first few characters of the first three or four words of the title, eight combinations in total.

(2) character strings consisting of 15, 25, 35, 45 or 55 characters from the beginning of the title, five combinations.

(3) same as (2) above, however, when the first word is an article, it is ignored.

3) Composition of the **PYAT** identifier (See **Table 2**)

Eight combinations of components of the **PYAT** identifier (**PYAT, PAT, PYT, PT, YAT, AT, YT,** and **T**) and variants of character strings, based on their length, extracted from each component produce a total of 144 kinds of identifiers. Of these, **A** has the first three characters of the last name of the author.

These identifiers are classified into three groups according to the type of the component **T** above as follows: W34 (i.e., 3 or 4 words) series which includes the **T** (1) above; TUM (i.e., title unmodified) series, **T** (2); and TAO (i.e., title article omitted) series, **T** (3).

In addition, the 145th identifier is a combination of **PYA**, of which **A** has the first five characters of the last name of the author.

4) Definition of "duplicates" in the **PYAT** experiment

Duplicates are defined as bibliographic items having the same **A, T,** and **Y** among all the components of the identifier. Although **P** is by definition not included as a component of the duplicate, it is included as a component of the identifier to measure its effect on identification efficiency.

5) Definition of identification effectiveness in the **PYAT** experiment

The success ratio in identifying duplicates in the population

$$= \frac{M}{D} \times 100 = \frac{M}{M+L} \times 100 = \alpha .$$

The success ratio in identifying duplicates by computer

$$= \frac{M}{C} \times 100 = \frac{M}{M+N} \times 100 = \beta .$$

M = The number of complete matches, i.e., the number of identical items among those identified by computer.
N = The number of noises, i.e., the number of wrong identifications among those identified by computer.
L = The number of omissions, i.e., the number of identical items which the computer failed to identify.
D = M + L = The number of real duplicates.
C = M + N = The number of items identified by computer.

The success ratio in identifying the duplicates in the population (α) and that in identifying the duplicates by computer (β) correspond to the recall ratio and the precision ratio in information retrieval efficiency, respectively. The number of real duplicates (D) is determined through manual checking prior to the identification experiment.

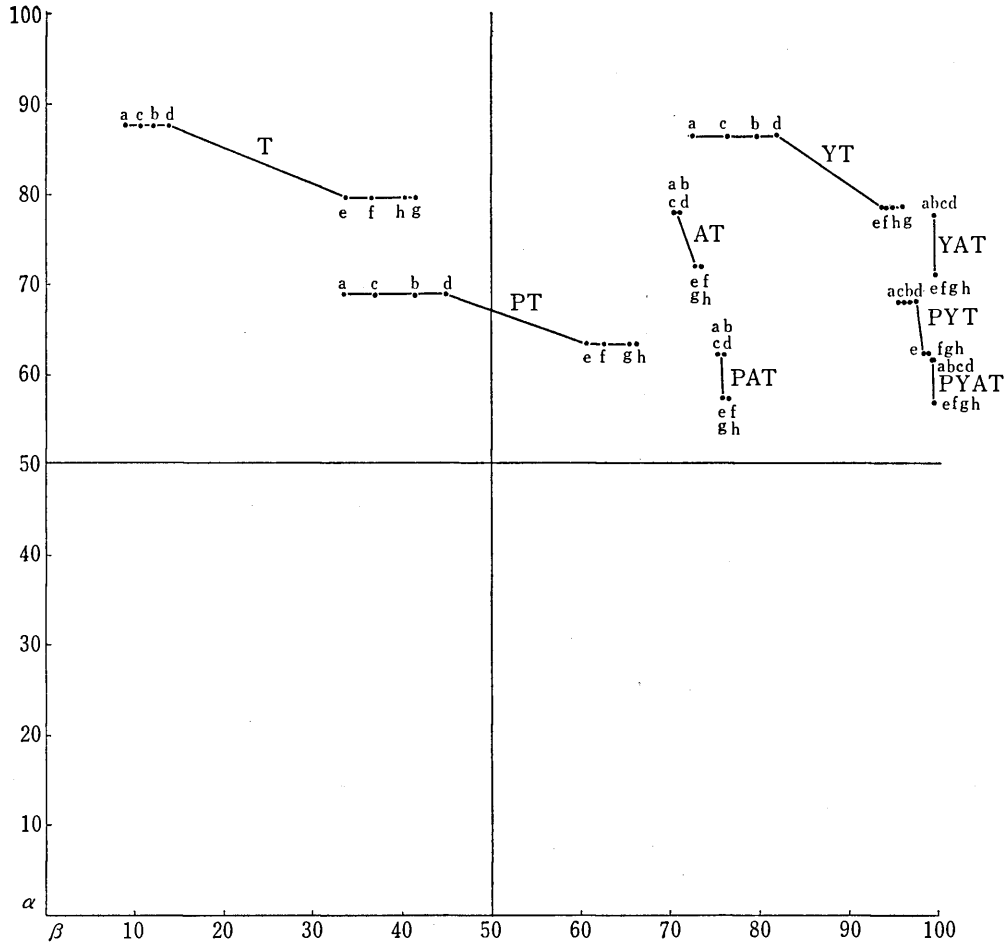
3. Results of the PYAT Experiment

In the **PYAT** experiment, the success ratios in identifying duplicates, α and β, are measured on the basis of manual checking of the results from computer identification. It was found that, when one component is eliminated from an identifier, α rises and β drops. This tendency is shown in **Fig. 1** (W34 series) and **Fig. 2** (TUM series). **Fig. 1** indicates that the identification efficiency is influenced more by the difference in the combination of the components than by the number of words extracted from a title. **Fig. 2** shows the changes in identification efficiency, according to the length of a title and the combination of components. The effect of each component, **P, Y, A** and **T**, is shown in **Table 3** in a TUM series of 35 characters.

Since **P** (place of publication) is by definition not included as a component of the identical items, the increase in α (e. g., 10.6% for **YAT**) is fairly large, and the decrease in β (e. g., 0.1% also for **YAT**) is small or almost negligible, when eliminating **P** from the **PYAT** identifiers.

Identifiers without **Y** (year of publication), such as **PAT, PT, AT** and **T**, show that the increase in α is small (e. g., 0.7% for **PAT**), while the decrease in β is very large (e. g., 19.7% also for **PAT**) compared with the combinations with **Y** (e. g., for **PYAT**). This means that books identified by combinations without **Y** include many different editions, and by definition are regarded as N.

Identifiers without **A** (Author), such as **PYT, PT, YT** and **T**, show that the increase in α (e. g., 5.3% for **YT**) is large, and the decrease in β (e. g., 0.5% also for **YT**) is small



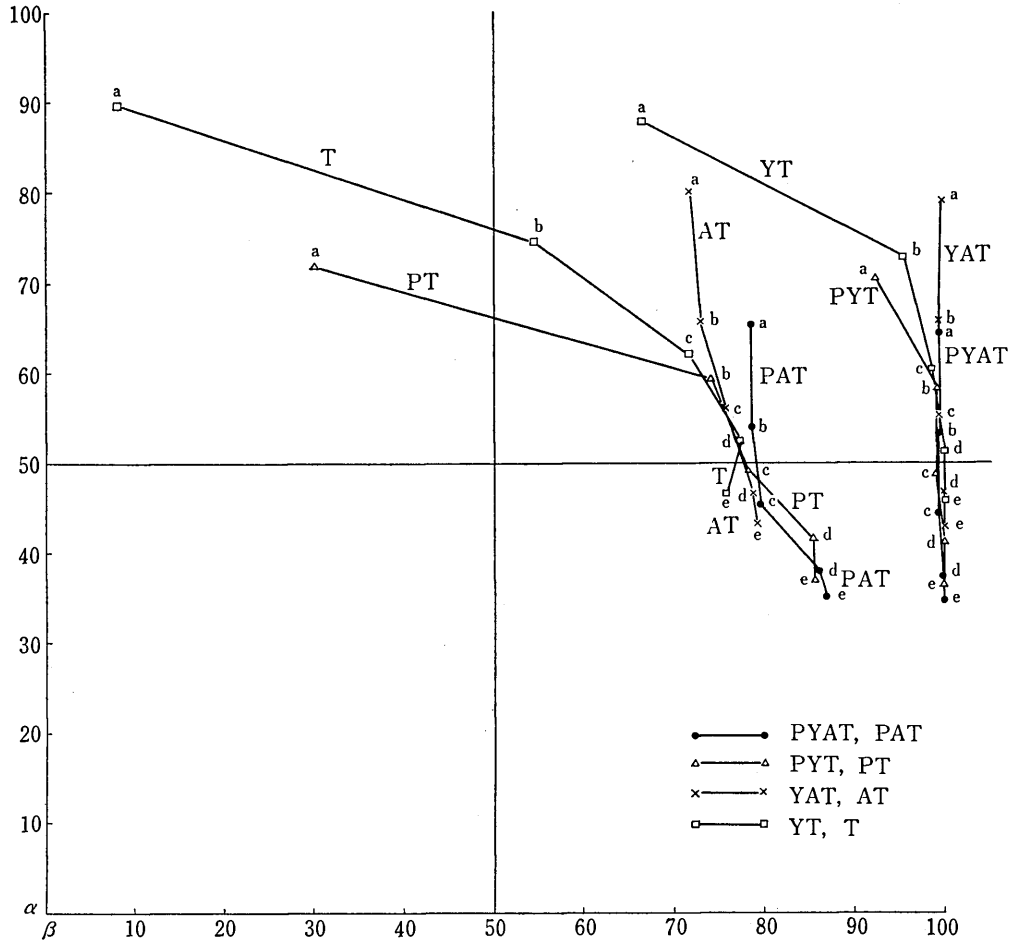
α =success ratio in identifying the duplicates in the population (recall ratio).
β =success ratio in identifying the duplicates by computer (precision ratio).

Fig. 1 Identification efficiency in W34 series of PYAT experiment

compared with the combinations having **A**. This indicates that description rules on author names are very different in the ECON and EBR files employed in the experiment. Moreover, author names actually appearing on the publications are in most cases described in abridged forms. This is one reason why an author's name is more effective when it is combined with other components for identification.

As for **T** (title), **PYAT** (with **T**) and **PYA** (without **T**) exhibit considerably different α and β values. That is, when eliminating **T**, the increase in α is 25.1%, while the decrease in β is 19.9%. This finding suggests that **T** is an indispensable component of an identifier.

Table 4 shows the causes of mismatches for components **P** and **A; Table 5**, for the



α =success ratio in identifying the duplicates in the population (recall ratio).
β =success ratio in identifying the duplicates by computer (precision ratio).

Fig. 2 Identification efficiency in TUM series of PYAT experiment

component **T**. In cases when mismatches are due to human mistakes, such as wrong spelling or different transliteration, accuracy of identification can be improved to a certain extent by devising automatic computer control. However, when mismatches are due to different description rules used in the two files, perfect automatic control is difficult to achieve, because the two description rules do not correspond to each other on an one-to-one basis. Yet, even then a **PYAT** identifier with a higher identification efficiency can be designed by thoroughly examining the description rules applied in advance to the two files.

4. Conclusion

In this study, the **PYAT** identifier was designed, and then an experiment was conducted

Table 3 Change of identification efficiency in TUM series of PYAT experiment*

component of identifier	change of combination	change of efficiency** α	β
P	P Y A T → Y A T	+10.6	−0.1
	P Y T → Y T	+12.6	−0.4
	P A T → A T	+10.3	−4.1
	P T → T	+12.2	−6.6
Y	P Y A T → P A T	+0.7	−19.7
	P Y T → P T	+1.1	−20.8
	Y A T → A T	+0.4	−23.9
	Y T → T	+1.2	−27.0
A	P Y A T → P Y T	+3.8	0.0
	P A T → P T	+4.2	−1.1
	Y A T → Y T	+5.3	−0.5
	A T → T	+6.1	−3.6
T	P Y A T → P Y A	+25.1	−19.9

*TUM series of 35 characters (see **Table 2**).
** α = success ratio in identifying the duplicates in the population (recall ratio).
β = success ratio in identifying the duplicates by computer (precision ratio).

Table 4 Causes of mismatches between ECON and EBR files in the cases of components P and A

causes of mismatches	component P	component A: personal author	component A: corporate author	component A: personal/corporate	component A: total
No description in EBR file	27	1	2	0	3
Differences in description rules	1*	2	3	0	5
Publication places different	24	—	—	—	—
Differences in authorship statement	—	4**	0	4***	8
Misspelling in either file	0	8	0	0	8
Total	52	15	5	4	24

*"Hague" in EBR file, but "The Hague" in WLS file.
**Editor in EBR file, but author (i.e., writer) in ECON file.
***Personal author in EBR file, but corporate author in ECON file.

Table 5 Causes of mismatches between ECON and EBR files in the cases of component T

causes of mismatches	types of identifier: W34, TUM & TAO	W34 & TAO	TUM & TAO	W34	TUM	TAO
Short title in ECON	4 (E2, F1, G1)*			8 (E5, F2, R1)		
Omission of subtitle in ECON				6 (E5, F1)		1 (E1)
Omission of the first article in ECON					2 (E2)	
Differences in description rules**	2 (E2)			1 (E1)		
Differences in punctuation rules***			2 (E2)			
Differences in spelling rules****	2 (E1, F1)	2 (E1, F1)		1 (G1)		1 (R1)
Differences in transliteration rules			2 (G1, R1)			
Misspelling in either file	2 (E2)		2 (E1, G1)			
Total	10 (E7, F2, G1)	2 (E1, F1)	6 (E3, G2, R1)	16 (E11, F3, G1, R1)	2 (E2)	2 (E1, R1)

*Initials in parentheses are language (e. g., E=English, 2=number of co-occurrences).
**For example, "Jan Tinbergen selected papers" in EBR file, but "Selected papers…" in ECON file.
***No space after comma in EBR file, but one space in ECON file.
****For example, "Produktionstheorie" in EBR file, but "Produktions-theorie" in WLS file.

to identify duplicates contained in two machine-readable files produced at different institutions in different periods. The identification experiment clarified that components of identifiers give different effects on the identification efficiency. Therefore, the identification efficiency thus acquired can be evaluated as useful data for composing a more effective identifier. To construct bibliographic databases, ways for efficient identification should be sought. Cooperative construction of databases has been initiated and it achieves considerable success only with such ways.

Acknowledgments

The ECON file used for our experiment was kindly offered by Mr. Charles W. Husbands, former Systems Librarian at Harvard College Library.

References

1) Matsui, S., Asano, Ch. and Matsuda, Y. (1989). A system for generating publication statistics based on bibliographic information. *Hitotsubashi Journal of Economics*, **30**, (2), 121-156.

2) Matsuda, Y. and Matsui, S. (1983). Shakaikagaku bunken no sokyuteki kensakuyou detabesu no hensei [Compilation of the database for retrospective search]. In: Kose, T. and Matsuda, Y., eds. *Shakaikagaku Bunya niokeru Deta no Detakozo no Bunseki* [*Analysis of the Data Structure of the Data in the Field of Social Sciences*], Tsukuba, RINET Society, 113-129.

3) Matsui, S. (1977). Shoshijoho detabesu no togo nituite [Techniques of identifying the duplicate items to integrate bibliographic data bases]. *Toshokan Tanki Daigaku Kiyo*, **14**, 113-139.

4) Harvard University Library. (1970). *Widener Library Shelflists, Nos.* 23-24: *Economics and Economic Periodicals*, Cambridge, Mass., Harvard University Press.

5) Matsuda, Y. and Matsui, S., comp. (1973). *Bibliography: Works on Econometrics — Reviewed in Econometrica, 1951—1971*, Otaru, RINET Society.