The ideas expressed in this paper are those of the author. The paper does not necessarily represent the views of the Center for Naval Analyses.
ERROR DETECTION IN COMPUTERIZED INFORMATION RETRIEVAL DATA BASES

Russell C. Coile

Center for Naval Analyses
University of Rochester
Arlington, Virginia

Centre for Information Science
The City University
London, England

Paper to be presented at the Sixth Cranfield International Conference on Mechanized Information Storage and Retrieval Systems, Cranfield Institute of Technology, Cranfield, Bedford, England

26-29 July 1977
ERROR DETECTION IN COMPUTERIZED INFORMATION RETRIEVAL DATA BASES

Russell C. Coile

ABSTRACT

The introduction of on-line interactive literature searching systems in recent years has made it possible for information scientists to conduct bibliometric studies which might have been difficult or impractical to do by manual methods. The unconventional uses of on-line information retrieval systems are becoming more common as we learn how to search using non-subject information fields. Author's name, organizational affiliation, journal's name, year of publication, etc., can now be searched for easily.

However, sometimes there are problems. For example, if the name of the author in a data base such as MEDLINE is given with initials for first and middle names, Bloggs, J. B. may be confused with Bloggs, J. B. since Joseph Blackwell Bloggs may be a mathematician while James Blackwood Bloggs is a chemist.

It would seem worthwhile for those responsible for management of these mechanized information storage and retrieval data bases to attempt to use all economically feasible error-detecting and correcting schemes to reduce the error rate as much as practicable. Several suggestions for detecting errors have been examined.
INTRODUCTION

Mechanized information storage and retrieval systems have brought a new era to information science and library operations. However, along with added flexibility and speed of searching and retrieval, we have become faced with more stringent requirements for accuracy in data bases. Unconventional uses of on-line information retrieval systems are becoming increasingly common as we learn how to search using non-subject information fields. We can now search easily using author's name, organizational affiliation, journal's name, and year of publication.

However, the increasing volume of scientific and technical literature has provided an impetus for more automatic error detection procedures to supplement the traditional human error detection and correction routines. The question to be considered is whether or not the combination of human and computer error detection systems is now able to cope with the volume of scientific literature.

Errors in Data Bases

For purposes of discussion, some of the illustrations of types of errors will be drawn from the Science Citation Index. This should not be misconstrued as being an attempt to publicize any presumed shortcomings of this data base. On the contrary, the Institute for Scientific Information has already taken extraordinary steps to correct errors in its data bases. As Sher (1) pointed out in a symposium on error control in chemical literature
during a meeting of the American Chemical Society in 1966, the data found in Index Chemicus are sometimes more accurate than in the original article from which the abstract was prepared. The error detecting procedures apparently included recalculation of molecular formulas by chemical abstracters who then requested the original author to confirm corrected errors.

We must also keep in mind that there may be different orders of importance of errors. Dr. Cawkell (2) classified errors in Science Citation Index into two major classes. A class one error would be one in which the result is that an item is very unlikely to be retrieved in consequence. A radical misspelling of an author's name might be an example of a class one error. A class two error would be of the kind which will usually not result in retrieval loss. For example, a non-standard abbreviation of a journal title might be a class two error since the cited item would appear beneath the correct cited author, usually in juxtaposition to the same item correctly cited (always assuming that the item has been cited more than once).

Errors in Primary Literature

Some errors originate with the author. For example, an erroneous reference or mathematical error will be published if it is not noticed by the referees and the editor. When the error is subsequently detected, an erratum may be published. If a reader detects the author's error, a reader's letter to the editor may
be published. The Science Citation Index serves a useful function in alerting people to this method of error correction just by tying together the later letter to the editor with the correction to the original publication which contained the error.

Even if the author were correct, errors might creep in through typographical misprints. An example of a minor typographical error is illustrated by a point raised in a letter to the editor (3) which commented that a paper had listed Alfred J. Lotka with the wrong middle initial of "K" in the first reference although correctly as "J" in another reference. Unfortunately, the Institute for Scientific Information's automatic error detection and correction program which will correct the misspelling of an author's name didn't catch the wrong initial, and the initial listing in the Jul - Sep 1974 LAHI to Z Science Citation Index repeated the error. The subsequent 1970-74 summary compilation corrected this error.

Another minor typographical error which slipped past the automatic error detection and correction program for misspelled author's name is illustrated by Droop's entry in the 1970-74 SCI for Lotke, A. J.

25 Elements Physical Bi
Droop MR Am Zoolog 13 209 73

There were actually 49 citations earlier to Lotka, A. J.

25 Elements Physical Bi

and in theory, the computer should have noticed the misspelling of
Several other examples of errors are discussed in another letter to the editor (4). The author's name of a reference had been misspelled, i.e. Learnes was listed instead of the correct spelling of Leavens. The name of the journal, *Econometrica*, and the year, 1953, were correct in this case. A more serious error, perhaps, was the statement in the paper that a particular 1941 reference showed that the number of authors fit a Yule-type distribution. First, this reference was the wrong reference since it didn't discuss the Yule-type distribution. Second, the correct reference, which was not given, should have been to a paper by Simon (5) in 1955 who had examined a probability model developed in 1924 by Yule (6) in connexion with analysis of the distribution of biological genera by number of species. Simon had proposed the application of this Beta-function model to frequency distributions of scientific publications, calling it the "Yule" distribution. In this case, having the wrong reference is probably a less important error than error of omission of the correct reference. Looking up the wrong reference may be a waste of time and frustrating, but not being able to consult the correct reference might waste a good deal more time in Sherlock Holmes type activity to find it.

Science Citation Index

There are several potential problems for a citation index. The first of these is the question of the cited author's name.
The author might change the way he writes his name as author of a paper from time to time. He might be C.N. Parkinson on one publication, C. Northcote Parkinson on a second, Cyril Northcote Parkinson on a third, and (although I have not seen it) Cyril N. Parkinson on a fourth.

The author might complicate things by changing her name upon marriage. The author might change his or her name after emigrating. Derek John Price of the S. W. Essex Technical College wrote all of his 1946-1949 papers on infra-red emissity of metals at high temperatures, etc. as D. J. Price. Derek J. de Solla Price (7) on this side of the pond wrote that classic, *Little Science, Big Science* in 1963. However, Derek de Solla Price (8) is now the author's preference. The Library of Congress apparently disregards an author's preference and has an old-fashioned concept that consistency is a great virtue. All of the relevant catalog cards adjacent to the main reading room in Washington have been painstakingly altered to "Derek John de Solla Price". These include all the old Derek J. Price cards with John de Solla added as well as the newer Derek de Solla Price ones with John added.

Finally, let us suppose that the cited author is consistent for fifty years or more and always uses the same name, e.g. Joseph Blackwood Bloggs on all of his papers. Various citing authors may either: a) spell out his name in full, b) use initials, i.e. J. B. Bloggs, c) use combinations of spelling and initial, i.e., Joseph B. Bloggs, d) use less than complete names, i.e., Joseph
Bloggs, or e) use less than complete initials, i.e., J. Bloggs. And, of course, in addition to these variations of the citing authors, the editors of different journals may have different policies as to names in references.

The overall result of all of these variations is what might be expected. Table 1 illustrates the problem of one who has a longer name than the customary American-style John C. Doe. The citations to Little Science, Big Science over a ten year period give more of the so-called "Brownie Points" to Price, D.J.D. than to any of the other variants including the author's new preference for Price, DDS.

While on the subject of errors introduced by the citing authors, may I point out that although Little Science, Big Science was published in 1963, there are publication dates of 1965 in the 1968 SCI volume, of 1968 in the 1970 volume and of 1970 in the 1973 volume. Furthermore, some additional errors slipped through the system with entries for Big Science, Little Science published in 1963 in the 1967 and 1971 volumes as well as Big Science, Little Science 1964 in the 1973 SCI volume.

Suggestions

What can be done to improve error detection and correction procedures?

Would it be economically feasible to add a computer error-detection program that would sort and group together in some editing file all items with the identical cited paper (or book)
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PRICE D</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>PRICE DJ</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>PRICE DD</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>PRICE DJD</td>
<td>9</td>
<td>11</td>
<td>8</td>
<td>17</td>
<td>9</td>
<td>10</td>
<td>13</td>
<td>13</td>
<td>6</td>
<td>9</td>
<td>105</td>
</tr>
<tr>
<td>PRICE DJS</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>PRICE DDS</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>PRICE DL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|       | 12   | 14   | 13   | 19   | 14   | 18   | 16   | 22   | 12   | 14   | 154   |

Table I

Citations to Little Science, Big Science in SCI
particulars and identical cited author last name? For example, before everything for a year is loaded into the masterfile, if items were put into a working file where a printout of the group of all references to 63 Little Science, Big Science with cited author's name of Price were produced we would see something like this:

```
PRICE D
   63 LITTLE SCIENCE, BIG S

PRICE DDS
   63 LITTLE SCIENCE, BIG S

PRICE DJ
   63 LITTLE SCIENCE, BIG S

PRICE DJD
   63 LITTLE SCIENCE, BIG S

PRICE DJS
   63 LITTLE SCIENCE, BIG S
```

The error-detecting program would make the initial sorting on last name only, not using any initials. Then it could make comparisons of initials to see if the identical initials are present. Rules for correcting the erroneous initials could then be applied by a human editor to add, subtract or change initials to one standard identical set of initials for all the identical cited papers (or books). Correction of the author's initials might be based on a review of the cited document, or inspection of American Men of Science, or Who's Who, or previous year's SCI, etc. to determine what the cited author's first name and middle name(s) actually are. The first letter of the first name should obviously be used and then the first initial of each middle name should be used in
sequence up to the computer's limit of three initials. I would vote for using the first initial of each particle as if they are middle names. Thus Derek J. de Solla Price would be

PRICE, DJD

The computer could be reprogrammed for automatic inconsistency correction if the human editing was deemed too expensive. For example, the computer could select the most popular variant of initials. Price, DJD was the winner in each year from 1964 through 1973. Or the computer could check with data already in file for the previous year and be consistent from year to year. If the previous year had several variants this might be used to alert someone or the computer itself to keep on going back in time until it found a unique entry of initials and then make everything identical. One could almost argue that consistency, even if it were consistently wrong, would be preferable to sometimes right, sometimes wrong.

Most of the discussion above concerns the inconsistencies of various combinations of initials associated with the author of a particular cited document. I have only mentioned the problems of the document such as the various erroneous years of publication that were given to Little Science, Big S. nor have I considered the errors in titles such as Big Science, Little S. Of even greater importance is a big problem. How does one get all the cited documents credited to the true author? After a computer or human editing decision which decided that only Price, DJD was
indeed author of Little Science, Big S. how does one devise a system to get

PRICE DJ
47 P PHYS SOC 59 131

which was correct at the time and is still correct but inconsistent with the new PRICE DJD to be credited to Professor Price? And how does one get all papers by Price listed as PRICE DJD if the citing author only refers to him as PRICE D?

Journal Titles

Journal titles have problems similar to those of names of authors. A journal may change its name. For example the Journal of Terrestrial Magnetism and Atmospheric Electricity after 52 years of publishing suddenly became the Journal of Geophysical Research. After only three years, the Journal of the Operations Society of America became Operations Research. The Forestry Quarterly and the Proceedings of the Society of American Foresters united in a new Journal of Forestry which continued the volume numbers of the Forest Quarterly.

Citing authors may use different abreviations for journal titles or various editors may use different abreviated titles.

How do various data bases cope with journals with the same title e.g., Journal of Education published in Boston, Massachusetts and the Journal of Education published in London, England?

Can citing authors be depended upon to give the full title to journals to avoid confusing Library Science and Documentation published in New York with Library Science, with a slant to
Documentation published in Bangladore, India?

**Suggestion**

An interesting solution to some of the problems of error detection was mentioned by Addelston (9) at the symposium on error control in chemical literature. Dr. Fieser, author of *Topics in Organic Chemistry*, was quoted as follows:

"When a new book is prescribed for use in one of our courses, I offer a prize of $1.00 for each error discovered in order that the first reprinting can be corrected as fully as possible."

Perhaps some variation on this theme for our computerized data bases is in order.
REFERENCES


2. Caskell, A.E. Director of Research, Institute for Scientific Information. Personal communication.


6. Yule, G. Udny 1924. "A Mathematical Theory of Evolution, based on the conclusions of Dr. J. C. Willis, F.R.S". Philosophical Transactions B 1924; 213: 21-


CNA Professional Papers — 1973 to Present*

PP 103

PP 104

PP 105

PP 106

PP 107
Stoffel, Peter H., "Relating Factor Analytically Derived Measures to Exogenous Variables," 17 pp., Mar 1973, AD 75 S 820

PP 108

PP 109

PP 110

PP 111

PP 112
Gorten, Lawrence H., "EIS Atmospheric Noise Level Statistics for Project SANGUINE," 29 pp., Apr 1974, AD 76 S 669

PP 113
Gorten, Lawrence H., "Propagated Anomalies During Project SANGUINE Experiments," 5 pp., Apr 1974, AD 76 S 666

PP 114
Maloney, Arthur P., "Job Satisfaction and Job Turnover," 41 pp., Jul 1973, AD 76 S 410

PP 115

PP 116

PP 117

*University of Florida, "Research reported in part under Office of Naval Research Contract N00014-80-0273-0017

PP 118

PP 119
Stoffel, Peter and Lockman, Robert F., "Development of Navy Human Relations Questionnaire," 2 pp., May 1974, (Published in American Psychological Association Proceedings, 81st Annual Convention, 1972) AD 77 S 246

PP 120

Economic, North Carolina State University.

PP 121

PP 122

PP 123

PP 124

PP 125

PP 126 — Classified.

PP 127

PP 128

PP 129

PP 130

PP 131

PP 132

Economic, North Carolina State University.

PP 133
Squires, Michael L., "A Stochastic Model of Regime Change in Latin America," 42 pp., Feb 1975, AD 79 S 912

PP 134

PP 135

PP 136

PP 137

* CNA Professional Papers with an AD number may be obtained from the National Technical Information Service, U.S. Department of Commerce, Springfield, Virginia 22151. Other papers are available from the author at the Center for Naval Analyses, 1401 Wilson Boulevard, Arlington, Virginia 22209.
CNA Professional Papers – 1973 to Present (Continued)

PP 174

PP 175

PP 176
Felix, Wendi, "Correlates of Retention and Promotion for USNA Graduates," 38 pp., Mar 1977

PP 177
Locke, W., Robert F. and Warner, John T., "Predicting Attitude: A Test of Alternative Approaches," 33 pp., Mar 1977. (Presented at the UN/DON Conference on Enlisted Attitude Xenia International Training Center, Xenia, Virginia, 4-7 April 1977)

PP 178

PP 179

PP 180

PP 181

PP 182

PP 183
Kraus, David, "Changes in Soviet Naval Forces," 33 pp., November 1976. (To be published as a chapter in a book published by The National Strategic Information Center)

PP 184

PP 185
Kraus, David, "New Technology and Naval Forces in the South Atlantic," 22 pp. (This paper was the basis for a presentation made at the Institute for Foreign Policy Analyses, Cambridge, Mass., 28 April 1977.)

PP 186

PP 187

PP 188

PP 189

PP 190