

Current Comments®

Quality Control at ISI: A Piece of Your Mind Can Help Us in Our Quest for Error-Free Bibliographic Information

Number 19

May 9, 1983

For several years, I've invited *Current Contents*® (CC®) subscribers and other ISI® customers to give me a "piece of your mind" about our products and services. We've regularly mailed questionnaires to a sampling of CC subscribers, and asked them to comment on any of the ISI services they use—how useful they are, how they might be improved, and what we are doing right or wrong. Figure 1 shows an example of a "Piece-of-Your-Mind" questionnaire. Your feedback helps us to maintain and improve the quality of our products.

Analysis of the responses shows that most people who use our services are satisfied. But about 25 percent of the responses include complaints about one thing or another. Most of them simply request a change in their address. While it is understandable that there is mobility among scientists and scholars, it is remarkable how often these changes occur. But it would seem that the actual user of CC rarely sees the envelope in which it is mailed. On the back of the envelope is a clear reminder to correct any changes in address.

Another popular complaint concerns journal coverage. Usually it involves a particular journal that is not covered in the CC edition the subscriber uses. Quite often it is covered in another ISI product. These recommendations are forwarded to Susan Jones or other members of the journal evaluation department. Readers are usually surprised to learn how carefully we appraise each journal recommended. As any CC reader

knows, we must strictly limit the number of journals we can cover.

In addition to soliciting comments from subscribers, there is a more direct method of quality control employed at ISI. Each week we write to a sample of 100 authors whose articles have been indexed in our data bases. All authors receive a computer printout of the complete bibliographic information we have extracted from their articles. This proof-reading procedure turns up a small number of errors. But occasionally those may be symptomatic of a more general problem, which we promptly investigate.

Some of the complaints I receive directly are from authors who have spotted an error in one of our indexes. Although we have devised a rigorous system to control the quality of our data, a small fraction of errors still avoids our detection. In most cases, the mistake is traced back to the primary source of our data—the scholarly and technical journals and books we cover. Nevertheless, we still get the "bum rap" for these errors.¹

ISI faces an extraordinary quality control challenge. We have to process an incredible amount of bibliographic information and make it available in a very short time. Over 2,000 source articles containing 30,000 references are processed *per day*. The data we extract from these publications have to be keyed into our computers, verified and corrected when necessary, and sorted at the end of each week for CC, *Automatic Subject*

Figure 1: Sample of a "Piece-of-Your-Mind" questionnaire that is mailed to ISI® subscribers.

Please check the ISI services you use and comment on them in the space provided below. If you would like more information on any ISI service, please circle the appropriate box.

Current Contents®

- Life Sciences
- Physical, Chemical & Earth Sciences
- Agricultural, Biology & Environmental Sciences
- Clinical Practice
- Social & Behavioral Sciences
- Engineering, Technology & Applied Sciences
- Arts & Humanities
- CompuMath®
- GeoSciTech™
- Quarterly Index to Current Contents®/Life Sciences™ (QUICC™/LS)
- Science Citation Index® (SCI®)
- Social Sciences Citation Index® (SSCI®)
- Arts & Humanities Citation Index™ (A&HCI™)
- CompuMath Citation Index™ (CMCI™)
- GeoSciTech Citation Index™
- Current Bibliographic Directory of the Arts & Sciences (CBD®) (formerly ISI's Who is Publishing in Science®)
- Index to Scientific Reviews™ (ISR™)
- Index to Scientific & Technical Proceedings® (ISTP®)
- Index to Social Sciences & Humanities Proceedings® (ISSHP®)
- ASCA® (Automatic Subject Citation Alert)
- ASCATOPICS®
- ISI® Search Service
- ISI Atlas of Science™: Biochemistry and Molecular Biology 1978/80
- ISI Atlas of Science™: Biotechnology and Molecular Genetics 1982
- Request-A-Print™
- Original Article Text Service (OATS®)
- ISI Press® Publications
- Educational Lecture Program
- ISI Catalog of Services
- Current Controversy

Chemical Information Division (CID)

- Index Chemicus Registry System® (ICRS®)
- Current Abstracts of Chemistry and Index Chemicus® (CAC&IC®)
- Chemical Substructure Index® (CSI®)
- Current Chemical Reactions® (CCR®)
- Current Abstracts of Chemistry and Index Chemicus® Microform Cumulation
- Chemical Substructure Index® Microfilm Cumulation
- Automatic New Structure Alert® (ANSA®)
- Index Chemicus® Registry of Organic Compounds (ROC)

Online

- Via Dialog/BRS:
 - SCISEARCH® (Search Service for Science)
 - Social SCISEARCH® (Search Service for Social Science)
- Via ISI Search Network:
 - ISI/BIOMED™
 - ISI/ISTP&B™
 - ISI/CompuMath™
 - ISI/GeoSciTech™

Sci-Mate™ Microcomputer Software

- Universal Online Searcher
- Personal Data Manager

Tapes

- Customized Tape Service
- ISI/CompuMath™
- ISI/ISTP&B™
- GeoSciTech™
- Science Citation Index®
- Social Sciences Citation Index®
- Arts & Humanities Citation Index™
- Index Chemicus Registry System®

Comments Here:

Please print your name and phone number above if we may contact you to discuss your comments.

Please check the accuracy of your name and address on the label. Note any errors or changes directly on the label.

Citation Alert (ASCA®),² and our multidisciplinary online data bases, such as *ISI/CompuMath®*, *ISI/BIOMED®*, *ISI/GeoSciTech™*, and others.

The reason ISI has established a good reputation for the accuracy and timeliness of its products is due largely to the quality control procedures we've developed. This ensures that our data are not only timely but virtually error-free. As you will see, our computer-based methods can even detect and correct errors that were made in the original journals.

A common problem we face in processing journal references involves the spelling of a cited author's surname. The basic function of our citation indexes is to enable users to retrieve current publications that have cited a particular author or article they are interested in. If a cited author's name is misspelled in a reference, the citing article may not turn up in a search of *Science Citation Index® (SCI®)*.

Anyone who submits an article for publication should be responsible for the accuracy of every cited reference. Journal editors share this responsibility when the article is accepted. Spelling errors would never get beyond the galleys if everyone did his or her job. But authors and editors aren't always as careful as they should be. So a lot of errors make their way into the journals we process. While the number is small in relation to the number of entries we process, the burden of detecting mistakes is not trivial.

One way we can detect errors is to compare the cited reference against the entries we create for the original article. ISI's files go back to 1955 and now include information on nearly ten million articles. However, we've stored complete bibliographic information on every source article we've processed since 1970. This "Forever File" includes accurate records on more than five million articles. Before we prepare our annual citation indexes and cumulations for publication, all the citations we processed during the year are run against

the Forever File to check for errors. About 47 percent of the citations are to articles in this Forever File.

In one of the algorithms used, the computer reduces the already condensed citation to a 14-character code. The abbreviated code depends on the first four letters of the author's name, the year of publication, and the volume and page of the journal in which it appeared. The computer then searches the Forever File for the source article that matches this abbreviated citation. When a complex computer algorithm confirms that a match is found, the computer "rewrites" the citation to conform with the bibliographic data on the source article in the Forever File.

Figure 2 shows a selection of entries from a printout generated after the 1980 *SCI* citation file was matched against the Forever File. In each case, the first line is the erroneous citation extracted from the original article. The second is the correct citation obtained from the Forever File.

For example, in a paper published in 1980 we processed a reference to A. Agostini's article in *Lancet* 1:453, 1978. The 14-character code for this article is AGOS78 1 453. The Forever File contained a matching record but the author's name had been misspelled in the reference. The computer "corrects" this error by replacing the citation with the bibliographic data from the source article.

As you can see, surnames and/or initials can be miscited in a variety of ways. Some are simply misspelled—CW Bank versus CN Banks. Many aren't even close to the correct spelling—BE Chattman instead of BE Chatterton, and D Flangman versus DR Flanagan. Others are incomplete—PS Weathers instead of PS Weathersbee.

If an author has a compound surname, many variants of it may be cited. For example, K Lindahl-Kiessling was cited as KK Lindahl—half of the surname was carried over as an initial. Compound surnames are particularly

Figure 2: Sample from a printout of ISI®'s existing program to detect errors in the spelling of a cited author's name. In each pair, the first line is the citation as it appeared in an article's references. The second line shows how the citation is automatically abbreviated and matched against the "Forever File" for the correct spelling of the cited author's name.

AGOS78	1	453	AGOSTINI A AGOSTONI A	LANCET LANCET	1	453	78
BANK67	51	566	BANK CW BANKS CN	BRIT J OPHTHALMOL BRIT J OPHTHALMOL	51	566	67
BAR581	6	418	BARSOTLE A BARSOTTI A	CLIN NUCL MED CLIN NUCL MED	6	418	81
BON179	75	565	BONILLA M BONILLASIMON M	AN QUIM AN QUIM	75	565	79
CHAT81	54	1116	CHATTMAN BE CHATTERTON BE	BRIT J RADIOL BRIT J RADIOL	54	1116	81
CHEN71	272	955	CHENOCULT J CHENAULT J	COMP REND C C R ACAD SCI C CHIM	272	955	71
FLAN76	41	3118	FLANGMAN D FLANAGAN DR	J ORG CHEM J ORG CHEM	41	3118	76
FRAU64	1	173	FRAUMARI JF FRAUMENT JF	LANCET LANCET	1	173	64
LIND76	19	1365	LINDAHL KK LINDAHLKIESSLING K	LIFE SCI LIFE SCI	19	1365	76
WEAT75	43	141	WEATHERS PS WEATHERSBEE PS	J REPROD FERT J REPROD FERT	43	141	75

troublesome because there are no universal standards. I've discussed ISI's policies on indexing compound surnames in a separate essay.³

The procedure I've just described, however, assumes that most people get the first four letters of the last name spelled correctly. But this assumption isn't always valid. In fact, we sometimes find that the wrong name has been cited. In the past few years, we've developed a new procedure that will detect spelling errors in any part of the cited author's name. Let me add that this algorithm won't replace the unification procedure described above. After the new algorithm is tested successfully, it will be added to our existing quality control procedures to enhance the overall accuracy of our data.

The new algorithm is not dependent on the accuracy of the first four letters of the cited author's name. In this case, we sort the files in alphabetical order by pre-unified *journal* rather than by author. If the variant author's name on a cited article can be matched with the author of an article from our Forever File, the cited variant is changed. If no

match is found, all the variant spellings of an author's name for the cited journal article are analyzed, and the computer makes a human-like decision as to whether or not they are the "same" author. After this decision is made, the computer chooses the "winner"—the citation that is most probably correct.

To do this, the computer keeps track of the number of times each variant was cited. Based on these data, the algorithm defines two requirements for the winning citation: it must have at least *three* citations, and it has to have at least *twice* as many citations as any other variant. If a citation meets these requirements, then all other variants are unified with it. That is, all the variants are corrected to conform with the winning citation.

Figure 3 shows a sample from a printout we obtained when the 1980 *SCI* citation file was run against this new program. The first two lines are an example of an article cited under two variants of the author's compound surname. This 1971 article, by P. DeMayo, was published in volume 4 on page 41 of *Accounts of Chemical Research*. In 1980, it was cited seven times under DEMAYO

P and eight times under MAYO PD. The computer would leave these citations as they are for the moment, because there is no clear "winner"—the most-cited version wasn't cited at least twice as often as the other variant.

The next two lines are another example of an article cited under different versions of an author's compound surname. The article was cited once under MACMURRY JE and nine times under MCMURRY JE. The computer recognizes the latter variant as the correct citation—it received more than three citations, and was cited more than twice as often as the other variant. The computer would automatically unify all the citations to this article with MCMURRY JE.

The next three lines show that an article was cited under three different spellings of the author's name. In eight cases REIKE RD is listed as the author, whereas the other two spellings are used only once each. So, the computer would select REIKE RD as the preferred citation.

The last two lines are an interesting example—two completely different authors were cited for the "same" article. Nine citations listed HANSMA PK as the author, and one listed KIRTLEY J. As it turns out, Hansma was the primary author of this paper and Kirtley was the secondary author. But the computer would *not* unify these "unmatched" citations under HANSMA PK, even though

it looks like a winner. The reason is simple—several letters, abstracts, or other short communications can appear on the same page of a journal. It would be a mistake on our part to unify them all.

I should stress that we are still testing this new method of detecting and correcting misspellings of cited authors' names in the journals we process. We may find that the number of variants without a clear winner is too small to justify spending a lot of expensive computer time to match listings in a file of millions of records. We have to balance the costs against the benefits our users might derive from it. However, there are now alternative procedures we can adopt due to changes in our computer system architecture. These new procedures may also make it possible to do these corrections in real time, that is, as soon as they are keyed into the computer. They would be integrated with the *Keysave* system I have described previously.⁴ The *Keysave* system detects many errors at the moment our data entry staff keys the citations.

As you can see, it takes a considerable effort on our part to correct references that misspell or misrepresent a cited author's name. But that's only half of the problem—other parts of the reference may also be incorrect or incomplete. For example, the citing author may leave out the volume or page numbers in a reference. Or the wrong year or volume may

Figure 3: Sample from a printout of ISI's new computer program to detect errors in the spelling of a cited author's name. The computer lists all variant spellings next to the record for the cited article, and the number of citations each received. Citation data are used to identify the correct version, and the computer unifies all variants with the "winner."

Bibliographic Data						Citations
ACCOUNT CHEM RES	1971	4	41	DEMAYO P		7
ACCOUNT CHEM RES	1971	4	41	MAYO PD		8
ACCOUNT CHEM RES	1974	7	281	MACMURRY JE		1
ACCOUNT CHEM RES	1974	7	281	MCMURRY JE		9
ACCOUNT CHEM RES	1977	10	301	REICKE R		1
ACCOUNT CHEM RES	1977	10	301	RICKE RD		1
ACCOUNT CHEM RES	1977	10	301	REIKE RD		8
ACCOUNT CHEM RES	1978	11	440	HANSMA PK		9
ACCOUNT CHEM RES	1978	11	440	KIRTLEY J		1

be cited. Citations to articles in journal supplements complicate the problem even more. The supplement number may be miscited as though it were the volume or page, or vice versa.

We have programmed, and are now using, an algorithm designed to detect and correct errors in the volume or year of cited journal references. The algorithm instructs the computer to first sort the file of citations in alphabetical order by the cited author. The computer then lists all the articles by the cited author that began on the same page of a given journal. If the wrong volume or year was cited, all the variants will be shown, including the number of times each was cited.

The computer automatically unifies the variants under a single citation if it recognizes a winner. Again, the winning variant must meet citation count requirements similar to those I described above. If there is a clear winner, and a variant differs on *both* the volume and year, the computer will not unify them. The variant must agree with the winner on one or the other—volume or year—before it is unified.

Figure 4 shows an example from a printout generated after the 1980 *SCI* citation file was run against this program. The correct citation is indicated by the symbol "***W***". As you can see, R.R. MacGregor's 1974 article in the *New England Journal of Medicine* was cited correctly 37 times. But one citation

listed the wrong volume number—192 instead of 291. Two more citations incorrectly listed volume 271. And one citation got the year wrong—1976 instead of 1974. The computer automatically corrects and unifies these variants under the winning citation.

The same types of errors were made when R.M. MacLeod's 1974 article was cited. In one case, a citation failed to include the year. And five citations to O.P. Mehra's article didn't include the volume number of the cited journal.

The last line in Figure 4 shows a variant citation that disagrees with the winner on *both* volume and year. In this case, the computer indicates that it didn't match on either, and would not unify this variant with the winning citation.

Figure 5 shows an example of an article published in a journal supplement that was cited many different ways in 1980. A. Böyum's 1968 article in the *Scandinavian Journal of Clinical and Laboratory Investigation*, volume 21, supplement 97, was cited correctly more than 530 times. But 94 citations listed the supplement number as the volume. Four citations did the opposite, and they failed to give *any* volume number. Six citations got the volume number wrong—20, 12, 22, and 24 instead of 21. And 13 citations listed the wrong year. Such problems in citation practices are a small indicator of the headaches created for librarians and others who must track

Figure 4: Sample from a printout of ISI's new program to detect errors in volume or year of a cited reference. The computer lists all variants next to the cited author's name, and the number of times each was cited. Citation data are used to identify the correct version, indicated by "***W***", and the computer unifies all variants with the "winner."

Bibliographic Data			Citations		
MACGREGOR RR	N ENGL J MED	291	642 74	*W*	37
MACGREGOR RR	N ENGL J MED	192	642 74		1
MACGREGOR RR	N ENGL J MED	271	642 74		2
MACGREGOR RR	N ENGL J MED	291	642 76		1
MACLEOD RM	ENDOCRINOLOGY	94	1077 74	*W*	93
MACLEOD RM	ENDOCRINOLOGY	25	1077 74		1
MACLEOD RM	ENDOCRINOLOGY	94	1077		1
MACLEOD RM	ENDOCRINOLOGY	94	1077 69		1
MEHRA OP	CLAYS CLAY MINER	7	317 60	*W*	21
MEHRA OP	CLAYS CLAY MINER		317 60		5
MEHRA OP	CLAYS CLAY MINER	7	317 65		1
MEHRA OP	CLAYS CLAY MINER	5	317 58		1 NO VOL/YEAR MATCH

Figure 5: Sample from a printout of ISI®'s new program to detect errors in volume or year of a cited reference to an article published in a journal supplement. The computer lists all variants next to the cited author's name, and the number of times each was cited. Citation data are used to identify the correct version, indicated by "*W*", and the computer unifies all variants with the "winner."

Bibliographic Data		Citations				
BOYUM A	SCAND J CLIN LAB S97	21	77	68	*W*	532
BOYUM A	SCAND J CLIN LAB INV	20	77	68		2
BOYUM A	SCAND J CLIN LAB INV	21	77	69		3
BOYUM A	SCAND J CLIN LAB I S	21	77	62		1
BOYUM A	SCAND J CLIN LAB I S	97	77	68		94
BOYUM A	SCAND J CLIN LAB S21		77	68		4
BOYUM A	SCAND J CLIN LAB S97	12	77	68		1
BOYUM A	SCAND J CLIN LAB S97	21	77	67		6
BOYUM A	SCAND J CLIN LAB S97	21	77	71		1
BOYUM A	SCAND J CLIN LAB S97	21	77	78		2
BOYUM A	SCAND J CLIN LAB S97	22	77	68		2
BOYUM A	SCAND J CLIN LAB S97	24	77	68		1

down these documents when they are requested.

Böyum's article was featured as a *Citation Classic* in *CC* last year.⁵ It has been cited over 6,000 times by now. On the basis of this article alone, Böyum should have been included in our list of 1,000 most-cited authors for work published from 1965 to 1978.⁶ Unfortunately, the file we used to do the study did not take into account books of any kind or journal supplements. In future studies, our procedures will improve citation counts to journal supplements.

These examples show why it is important for authors and editors to get their references straight before an article is published. We're doing everything we can to correct errors we find in the primary journals so that they are not repeated in our products. The procedures we have been developing have successfully detected thousands of misspelled names and other errors. Eventually, we will be able to correct errors or variations in pagination as well. Consider the practice of citing just the single page on which an observation is reported, rather than the first page of the article. By using our Forever File we will be able to unify these variations as well.

Even after all these computer-based quality control strategies become a part of our production system, I'm sure that a fraction of the errors in the primary

journals will still get into our data bases. Many references are wrong on *several* points—cited author's name, journal, volume, page, and/or year. They are so ambiguous that even a human can't decide what the correct citation should be. Our quality control algorithms have to be based on the assumption that most of the citation is correct. Otherwise, we would run the risk of unifying citations that should remain separate. That's why comments from authors are a necessary part of our quality control system. The people who use our products can reduce the fraction of errors in the primary journals that slip through our filters.

This year, our Piece-of-Your-Mind questionnaires will be mailed to *all* subscribers of ISI products. As in the past, I will read *every* reply—good, bad, or indifferent. Recently, I set up a minicomputer system to store all of these messages.⁷ Every letter of complaint is answered by me or handled by a member of the ISI staff.

Although I've concentrated on computer-based systems to detect errors, you should keep in mind that our staff is actively involved in ensuring the high quality of our products in other ways. Each department at ISI has its own quality control team. Our in-house quality control procedures help assure that *we* don't make mistakes in processing and entering bibliographic information in

our data bases. The information we process every day is verified on a character-by-character basis. If a data processor miskeys a reference, it is corrected *before* it is stored in our files. They do an excellent job. So it's disappointing when we get the "bum rap" for those errors that are the fault of editors and authors.

We also have a rigorous system to monitor the printing and binding quality of our products. Our cumulated sets of citation indexes have a very high inventory value, and they have to be *complete*. If a subscriber complains about a printing or binding flaw in any volume of an annual set of *SCI, Social Sciences Citation Index® (SSCI®)*, or *Arts & Humanities Citation Index™ (A&HCI™)*, we try to replace it. That breaks up a complete set in our stock that we can't sell. We simply can't afford errors in printing and binding, and we do our best to correct problems *before* a set is published. I could dwell at length on the quality control procedures that ISI has developed in printing and binding. Indeed, our procedures are now imitated by other firms that want to raise industry standards. A key person in this program is Irving Sher, ISI's director of quality control.

I want to stress that it's never too late to let us know about an error. Even if we can't correct mistakes that have already

been published, we can at least publish a suitable correction. And we can try to prevent repeating the error in our cumulated indexes or our online data bases. On this last point I should explain that there are, and must be, differences between some entries in our printed indexes and those found in online files. Until we are able to introduce our correction procedures on a daily or weekly basis, there will be differences due to the fact that online vendors mount our files as they are received each month. Once they are mounted at Lockheed, BRS, or DIMDI, it is not possible to incorporate changes we introduce by the large-scale editing procedures I have described to you.

ISI makes every reasonable effort to ensure the accuracy of the files. But we cannot guarantee it for every one of the millions of articles we process, either in our printed files or in the online files handled by other vendors. So keep your cards and letters coming in—if you have a complaint, you can be sure that I will read it and get back to you. As I said some time ago, "Learn to complain!"⁸ It will do us both some good.

* * * * *

My thanks to Alfred Welljams-Dorof for his help in the preparation of this essay.

© 1983 ISI

REFERENCES

1. **Garfield E.** Errors—theirs, ours and yours. *Essays of an information scientist*. Philadelphia: ISI Press, 1977. Vol. 2. p. 80-1.
(Reprinted from: *Current Contents* (25):5-6, 19 June 1974.)
2. You don't need an online computer to run SDI profiles offline! So why haven't you asked for ASCA—the ISI selective citation alert. *Current Contents* (13):5-12, 28 March 1983.
3. What's in a surname? *Current Contents* (7):5-9, 16 February 1981.
4. Project Keysave—ISI's new on-line system for keying citations corrects errors! *Essays of an information scientist*. Philadelphia: ISI Press, 1980. Vol. 3. p. 42-4.
(Reprinted from: *Current Contents* (7):5-7, 14 February 1977.)
5. **Böyum A.** Citation Classic. Commentary on *Scand. J. Clin. Lab. Invest.* 21(Suppl. 97):77-89, 1968. *Current Contents/Life Sciences* 25(45):20, 8 November 1982.
6. **Garfield E.** The 1,000 contemporary scientists most-cited 1965-1978. Part 1. The basic list and introduction. *Current Contents* (41):5-14, 12 October 1981.
7. Introducing *Sci-Mate*—a menu-driven microcomputer software package for online and offline information retrieval. Part 1. *The Sci-Mate Personal Data Manager*. *Current Contents* (12):5-12, 21 March 1983.
8. Learn to complain. The ultimate responsibility is with the individual, not the corporation. *Essays of an information scientist*. Philadelphia: ISI Press, 1977. Vol. 1. p. 465-6.
(Reprinted from: *Current Contents* (29):5-6, 18 July 1973.)