Current Comments

What's in a Surname?

Number 7 February 16, 1981

ISI® processes several million author names each year to produce the *Science Citation Index®* (*SCI®*), *Social Sciences Citation Index®* (*SSCI*™), and the *Arts & Humanities Citation Index*™ (*A&HCI*™). In 1979, we processed more than eight million authored source and cited items for the *SCI*. The half-million source articles alone involved over a million author postings. Although our quality control procedures are strict and thorough, we inevitably make some mistakes. But in most cases, the errors which users observe can be traced back to the original articles we process. But whether we or the citing authors commit the errors, they will stand out like a sore thumb if they turn up anywhere near the correct name.

Sometimes, but not always, these errors present an obvious obstacle to information retrieval.[1] The primary function of the Citation Index is to allow users to locate current works that have cited a specific article or author they already know. For this reason, it is especially important to index authors' names consistently. So when you are looking for an article by De Maggio, you should know whether it will be found under that heading or, as the case may be, under Maggio. *All* citations to the same author should be listed under a single, "standardized" surname heading. Otherwise, users would miss possibly relevant articles listed under variant spellings. These variants are the biggest source of our headaches in compiling our citation indexes.

Also, citation indexes may be used to compile citation counts of an individual's works for various evaluative purposes. For example, our highly-cited article studies are based on citation analysis. However, a given paper may not appear to be as highly cited as it really is because citations to it are listed under several variants of the author's name. Consider this example. Recently it came to our attention that citations to D. de Wied, State University of Utrecht, the Netherlands, appear in different volumes of the *SCI*! For example, in the 1979 *SCI*, his works were cited 328 times under DE WIED D, 186 times under WIED DD, and nine times under DE WEID D, the last being an outright misspelling! This illustrates that it is extremely important to sort all citations under a single surname to accurately reflect an individual's citation history.

People are generally sensitive both about the way their names are pronounced and how they are presented in print. Some authors may consider it a dishonor to their heritage when we abbreviate their names for our convenience. Many compound surnames indicate family roots, and they are usually prefixed by foreign articles and/or prepositions that translate into English as "of" or "from"—Wernher von Braun, John Dos Passos, George de la Tours, and Louis de Broglie, for example. Patronymic surnames indicate father or clan relationships—Douglas MacArthur, Gene McCarthy, and Pat O'Brien, for example. Also, hyphenated names usually signify the combination of two distinguished family lines, like Albert Szent-Györgyi.

Although compound surnames comprise only about five percent of all the names we process, there is a greater chance that variant forms of these names will occur than with other surnames. The basic problem is that oftentimes we can't tell what is the first, middle, or last name! This is particularly true for Oriental names. But it also applies to Romanian, Icelandic, and many Central European names. For example, Hungarians use the "last" name first when publishing in Hungarian journals. But the same authors writing abroad will put the "first" name first. In case you didn't know it, when in Iceland you would look for my friend Einar Sigurdson under Einar in the phone book. In the *SCI*, and in most of the world's phone books, he would be listed under Sigurdson.

Oriental names pose a special problem. For example, Chinese names list the surname first—Mao Tse Tung would be indexed under Mao, Tse Tung. But Chinese names containing a non-Chinese given name list the given name first—Philip Loh Fook Seng would be indexed as Loh, Philip Fook Seng.[2] The problem is compounded because there are only about 200 common Chinese surnames.[3] The same is true for Korean names. Although there are no more than 300 common Korean surnames, only *three* account for the great majority—Kim, Pak, and Yi![3]

Obviously, since a small class of common names accounts for the majority of Oriental surnames, *homographs* can be a serious problem. For example, the heading T SUZUKI in the *SCI* lists 945 papers, cited over 7,300 times from 1965 to 1978. Actually, there are more than 25 people named T SUZUKI in the *SCI*.

As I've pointed out before,[4] the homograph problem would be eliminated if authors used two or three initials in addition to their surnames. However, when using the *SCI* Source Index, homographs can be distinguished either by their addresses or the journals in which they publish. For example, the T SUZUKI of Akita University who published in *Experimental Parasitology* is not the same T SUZUKI of Sophia University, Tokyo, who published in the *Japanese Journal of Applied Physics.* We've included authors' addresses under the surname heading in the Source Index of the *SCI* since 1977. The journal has always been a part of both the Source Index entries as well as the Citation Index. Problems arise when trying to decide which of the many papers listed are by a particular author one is evaluating.

Muslim names also pose a problem—depending on the country of origin, there are *nine* different types of names a Muslim author may use, and the order of presentation varies from country to country. Anis Khurshid, director of the Islamic Library Information Center, University of Karachi, Pakistan, gives the following example of an Egyptian Arabic name: Fahr-ad-Din Abu Abdullah Muhammad Ibn Umer Ibn al-Hasan Al-Hatib Ar-Razi. As it turns out, Muhammad would be considered his "first" or given name, Ibn Umer Ibn al-Hasan his father's or forefather's name, and Al-Hatib his genealogical or tribal descent.[5] Fahr-ad-Din is an honorary title, Abu Abdullah is yet another name indicating descent, and Ar-Razi denotes the country or town of origin.

Also, the way a surname *should* be presented often depends on where the author resides. For example, a German author living in Germany would include the "von" prefix in the last name, but only if it is capitalized. If the "von" prefix is in lower case letters, it is usually included only as an initial. Another author with the same last name who lives in the US drops the "von" prefix, whether it is capitalized or not, or may adopt it as an initial. To add to the confusion, a Dutch or South African author would keep the "van" prefix with the last name, whether or not it is capitalized. Since it simplifies our computer procedures to use capital or upper case letters, we cannot honor such idiosyncratic conventions for Germanic or other surnames.

A very common source of error involves Spanish and Portuguese surnames. Although these surnames look very similar, there are different national conventions for indexing them. Portuguese names are indexed under the part of the name *following* the prefix.[2] Also, words like "junior" or "senior" are treated as part of the surname.[2] For example, Martinho Augusto da Fonseca Junior would be indexed as Fonseca Junior, Martinho Augusto da. On the other hand, Spanish names are indexed under the prefix itself if it is a single article, and words like junior are *not* included.[2] Thus, Manuel Antonio Las Heras Junior would appear as Las Heras, Manuel Antonio—but Antonio del Rio would appear as Rio, Antonio del because *del* is a preposition!

Even when we agree on the correct presentation of an individual's name, there is no guarantee that the name will be consistently spelled by citing authors. We tested the possibility of individualizing the way we handle names for my friend Derek John de Solla Price, Yale University historian of science. We instructed our personnel to index his name in a certain form, and to *watch* for his name in the future. We even wrote a special computer program to check whether his name was being indexed correctly. After all this effort, Price's name still was indexed incorrectly because it was presented in bizarre forms by citing authors.

Unfortunately, there is no convenient universal standard to which we can refer when indexing compound surnames. The closest thing to it is the *Anglo-American Cataloguing Rules*[2] (*AACR*), jointly revised in 1978 by the American Library Association, the British Library, the Canadian Committee on Cataloguing, the Library Association, and the Library of Congress. The *AACR* "has been adopted by major libraries and agencies in most English-speaking countries, and has had a considerable influence on the formation or revision of local and national cataloguing rules in a number of others."[2] The *AACR* includes more than 100 rules on how to index personal names, exhaustively covering compound names of European, Russian, Arabic, and Oriental origin.

The *AACR* was designed to help librarians decide what heading an author should appear under in a card catalogue. Significantly, the *AACR*'s general rule on choice of name reads as follows: "Choose, as the basis of the heading for a person, the name by which he or she is commonly known.... Determine the name by which a person is commonly known from the chief sources of information of works by that person issued in his or her language."[2] (p. 348) Librarians may be able to afford the time to search through the *AACR*'s 100 odd rules on personal name headings, and even to refer to sources from the author's country to determine the commonly used form of the name. However, ISI can't afford to spend time tracking down the "common form" of each compound surname we come across. A central feature of all our services is *timeliness*—our production schedules would be seriously delayed if we followed the *AACR*'s recommendations. And more importantly, the cost would be astronomical.

Even if we could follow their rules, we'd still receive complaints from authors. As I said above, some authors with foreign compound names no longer honor the conventions of their country of origin because they now live in the US. More seriously, some authors publish under various versions of their own surnames. They may include their entire family name on one article, or only a part of it on another. Or they may change the spelling of their surnames if the printer doesn't allow for accent or other diacritical marks—for example, Schröder *vs.* Schroeder. Lastly, articles are sometimes *cited* inconsistently—the first author of the original article may or may not be presented as the first author in the citation, or citing authors may abbreviate compound surnames according to their *own* rules!

It's just too complicated to treat each compound surname individually or to rely on standards like the *AACR*. ISI's Irv Sher, director of quality control, has evolved a more reasonable policy for consistently indexing these names. (I might add that this is based on 20 years of experience in dealing with the problem.) To begin with, an author's last name is tentatively defined as the *last* element (reading from the right) up to the first space we encounter. If the last element includes a hyphen, we'll accept the entire hyphenated name as the surname. If a "particle" (foreign articles, prepositions, and words indicating relations) immediately precedes it, we'll "collapse" it with the last name and accept the *sum* as the surname. For example, George de la Tours would be indexed as DELATOURS G. Table 1 lists those particles which will be collapsed with the author's last name. Of course, Oriental and Central European names that list the last name *first* will be handled separately. Most of this applies to our treatment of these names when they appear as the by-lines in original source articles.

Table 1: Particles commonly associated with compound surnames. These particles will now be combined with an author's last name, and the sum will become the heading under which that name will appear in the Citation and Source Indexes of the *Science Citation Index®*, *Social Sciences Citation Index®*, and *Arts & Humanities Citation Index™*.

AL
BEN
BIN
DA
DAS
DE
DE LA
DELA
DELLA
DEN
DI
DO
DOS
DU
D'
EL
IL
LA
LE

L'
MAC
MC
O'
SAINT
SAINTE
ST.
STE.
TEN
TER
v. (VON)
VAN
VAN DEN
VANDER
v.d. (VANDER)
VAN DER
VON
SEN
ZUM

But what do we do when names are spelled inconsistently by citing authors? How do we know that the 1976 article written by Paul De Maggio is the same article cited later by authors as P.D. Maggio? I've pointed out before that articles can be uniquely identified with a minimum of bibliographic information that doesn't include the author's full name.[6] In most cases, all you need to know is part of the journal title (or its abbreviation), volume, pagination, and year of publication—this is called a "condensed journal citation." No matter how the author's name is presented by various citing authors, the cited article can be identified as being the *same* article by using the condensed citation, except in rare cases when two articles or letters begin on the same page. (Adding the first letter of the author's last name eliminates any chance of error.) By using this coding algorithm, we can unify all references to the same source article. But how can we *correct* errors?

We are now preparing a file called the "Forever Dictionary." This file will include *every* source article we've processed for the *SCI* since 1961—over six million. The Forever Dictionary will also include an alternate record for any name prefixed by the particles listed in Table 1. When a 1976 article by Paul De Maggio is processed as a source item, we will store the alternate versions of the name. When the De Maggio article is subsequently cited as P. De Maggio, P. Maggio, or P.D. Maggio, these variant citations will be changed and later sorted together under the preferred DEMAGGIO form. That is, the "false" citations will be *corrected* to appear as if they referred to the standardized DEMAGGIO form. Whenever we correct a variant citation, the incorrect heading will still appear in the printed *SCI* as a "see reference" which will direct the user to the preferred form of the name. The Forever Dictionary will eventually include all source items processed for the *SSCI* and *A&HCI* as well. It may also include highly-cited papers used in our Keysave ™ system.[7]

In order to detect errors that defy the alphabetic look-up procedures outlined above, an alternative form of the Forever Dictionary is used. In this case, we sort the file by journal rather than by author. After matching on the condensed citation, we can match the standardized name with that actually used by the citing author. If we detect a difference, an editor can then make a "post-edit" decision to unify the variants. These correction procedures involve a great deal of work. ISI is cleaning up a lot of garbage dumped into the literature by careless authors and editors. Referees should insist that the spelling of cited authors' names be carefully checked.

I'm sure that a new set of authors may still complain about the way we now handle surnames under our new standards. We are trapped in a no win situation because we can't satisfy every author all the time. Our new rules aren't intended to dishonor an author whose name should ideally be presented according to the conventions of his or her country. Rather, they are intended to *honor* the author by bringing together all references to his or her works in a place where they can be found consistently in our indexes. We sometimes have to compromise between the wishes of the individual and the constraints of large indexes. As long as everyone is familiar with the rules, the compromise should work to the benefit of everyone involved. Of course, individuals who have encountered problems with the way their names are indexed are welcome to contact ISI to make sure we are indexing consistently.

It is not possible to correct errors in already printed indexes. But it is possible to correct machine readable records, and it is also possible to introduce corrections into our printed five-year cumulations. The next index of this kind for the *SCI* covers 1975-79, and will be published late in 1981. An enormous post-editing job has gone into cleaning up these files, not only for the reasons cited above, but also to eliminate "truncation," a practice dictated by our older computer methods but recently eliminated. All names are now completely spelled out in the Source Index of the *SCI.*

There is an old saying in Hollywood: I don't care what you say about me—just be sure to spell my name right. I think you can conclude that at ISI we're doing our best to preserve your name, no matter how other people may spell it.

* * * * *

My thanks to Patricia Heller and Alfred Welljams-Dorof for their help in the preparation of this essay.

REFERENCES

1. **Garfield E.** Errors—theirs, ours and yours. *Current Contents* (25):5-6, 19 June 1974.*
2. **Gorman M & Winkler P W,** eds. *Anglo-American cataloguing rules.* Chicago: American Library Association, 1978. 620 p.
3. **Zgusta L.** Names. *Encyclopedia Britannica.* Chicago: H.H. Benton, 1974. Vol. 12. p. 814-19.
4. **Garfield E.** British quest for uniqueness versus American egocentrism. *Nature* 223:763, 1969.*
5. **Khurshid A.** Is uniformity in cataloguing Muslim names feasible or possible? *Libri* 27:282-95, 1977.
6. **Garfield E.** Incomplete citations and other sources of bibliographic chaos. *Current Contents* (24):5, 17 June 1969.*
7. -------------. Project *Keysave*™ —ISI's new on-line system for keying citations corrects errors! *Current Contents* (7):5-7, 14 February 1977.*

*Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1980. 3 vols.