The Birth of AARON?

D. K. Tucker

Esson & Tucker Management Inc.

About 1455 AD the printing press was invented; two hundred years later the English language underwent spelling control. In another hundred years, in 1755, Samuel Johnson produced one of the first English dictionaries. Thus it took three hundred years from the invention of the first mass reproduction technology associated with language to the provision of a useful reference for language: the dictionary. Today, perhaps the most comprehensive book of its type is the *Oxford English Dictionary*; it contains over 500,000 words and the third edition is due to be published in 2010.

In 1951 not only was the American Name Society inaugurated, but Remington Rand introduced the world's first commercial computer: the UNIVAC (*UNIV*ersal *A*utomatic *C*omputer). The computer grew up and fell in love with the dictionary and now the OED and other dictionaries are available to any Internet user. Cousins of the dictionary sit on the millions of desktop computers, each of which has vastly more power than the UNIVAC. One cousin is a register of words, which drives the spellchecker, and another is the thesaurus. The words of our language are available; we know what they mean, where they came from, how to pronounce them, and how to replicate them in text without error. We also have a host of automated tools including grammatical and style advice, autocorrect, automatic typing from dictation, and suchlike. We may conclude that, although everything can be improved, it looks like the writer can be left to concentrate on the content.

Well, almost. I ran the MS-Word 97 spellchecker on this text and it balked at *Remington*. It didn't recognize it. *Remington*, a name that conjures up images of computers, typewriters, small arms, and cowboy art, halts the spellchecker, although it has no difficulty with *Microsoft*. *Remington,* for which the search engine Google produced "about 230,000 (hits) . . . in 0.31 seconds."

Names 49.4 (December 2001):240-244
ISSN:0027-7738

Type a business or personal letter, or an email, and run the spellchecker; it usually balks on the name and address bits like Mr. *Broadfoot* and *Manotick*. These are the bits you really should get right—the hygiene bits that get you nothing if you do get them right, but immediately lower the chance of achieving your objective if you get them wrong. These bits are usually called names, or more properly, proper names or proper nouns—the capital letter names.

Names like *Jane, Austen* (which the spellchecker wants to change to *Austin*) and *Highbury*. Names like *Jane Fairfax* and *Frank Churchill*. Names like *Doris Day* and *Marilyn Monroe*. Names like *Cadillac*, *De Ville*, and *General Motors*. Names like *Barbie, Pokemon* and *Sony*. Names like *Dodgers, Lakers, Celtics*, and *The Great One*. Names like *Survivor, Wall Street Week* and *Damage*. Names like *Google* and *Yahoo*. Forenames and surnames of people and characters; pseudonyms, stage names, place names; names of products and companies; names of sports teams; names of shows, songs, operas, films, tours, search engines, et cetera, et cetera.

Names give our culture meaning and structure but we do not have a ready reference to them. This is not so say that there are no references but they are anything but inclusive. How many names are there? I suspect that no one knows but we can get some idea. There are over 1.25 million unique forenames and 1.75 million unique surnames in the United States, so, ignoring overlap, we have three million to start with, then add the almost two million physical and cultural geographic features in the U.S. in the Geographic Names Information System (GNIS). (NIMA, the National Imagery and Mapping Agency, has almost four million entries of foreign geographic feature names.) For the U.S., therefore, we can readily identify five million entries; a conservative doubling and we get ten million—twenty times the number of entries in the OED.

Why should we want such a reference? It could be argued that the Internet is the reference source, but the search engines, which gather the information, have to have references to search and comprehensive lists of forenames and surnames are just not available. These ten million proper names permeate our culture and we know little about them, so the question is a little like asking why would we want a dictionary of all the words of our language. We have a huge cultural gap; over 95% of

the words available are these proper names. Names of places seem to be well looked after with the information available on the Internet. One can quickly establish from GNIS that there are 59 features in the U.S. with *Mulligan* in the name, where they are, and what type of feature they are. What we need is the same sort of thing for the other names.

There is a plethora of baby name dictionaries; although why we would call these "baby names" when our culture names for life is a puzzle. These dictionaries vary from excellent to less so, but most seem limited to about five thousand entries; often the same five thousand entries, it seems. There is not yet a comprehensive dictionary of surnames in the U.S. although Oxford University Press has long threatened to publish one. Perhaps a good place to start would be to find out more about *Mulligan*; that is, to concentrate on the forenames and surnames of the people in the U.S. and Canada.

The currently high level of interest in genealogy is readily ascertained by a few minutes browsing the Internet, and this concern has led The New York Times (August 12, 2001) to suggest that genealogical services are the second most visited sites on the Internet. The huge Mormon site reputedly experiences millions of hits per day. Another indicator is the success of the U.S. Census Bureau's site that lists the most popular forenames and surnames. As a matter of passing interest this site lists *Harold* in the list of female forenames; it is not a processing mistake, but could be a misreading of the census form by those who complete it. This is just one of the challenges to be faced in gathering such data. Away from the Internet, look at the popularity of "Map Your Family Tree" products in your local computer or bookstore.

The success of the OED has not prevented others from creating their own excellent dictionaries, nor has the success of GNIS prevented independent work in this area. In coming to grips with three million names, the first thing would be to generate the register from which others could create the dictionaries. A register differs from a dictionary in that the former contains no etymology; it is a list of names sometimes with frequency, or count, of the name. Such a register of forenames and surnames will not inhibit creativity but stimulate it; to state the facts and challenge others to explain them has long been a successful technique.

Is there any commercial benefit that would cause someone to invest in such a product as a names register? The answer is yes, but the need, although clear for all to see, is not obvious, as we have for a long time

accepted the situation. There are as many ways to program a computer to do a particular job as there are number of jobs to do: limitless. To improve productivity, organizations like to have programs that are tried, tested, and true (TTT) to execute particular functions and use that program whenever that function is required. Not only is this more efficient, it also avoids the problem of having a program that does not work under certain unusual circumstances. The TTT program is treated as a module much like a sub-assembly in manufacturing companies where the use of TTT modules is a cornerstone of modern production.

In text creation by keying we are allowed complete freedom to create rubbish like *kmdpD 39)"*. We might not spot it ourselves, but if we have it underlined in red for our attention we can always use the after-the-event process—spellchecker. Spellchecker says, "I don't recognize this," and may offer several alternatives. It works reasonably well and will no doubt get better.

So if I now want to write the name *Tchaikovsky* I get no help. The spelling used is an accepted transliteration but not everyone knows that. Incidentally, it is not just the surname of a dead Russian composer, but also the names of U.S. residents. Not being familiar with Russian names I might have misspelled it as *Tchiakovsky* and I would have been none the wiser. The keyboard is a marvelous device but it allows us to create rubbish; it allows us to skip out of the universe of known names without sanction. This might be OK if we are inventing names but not if we are trying to get the attention of a particular person. What we need is something like the current autocorrect function. We need the spellchecker register enlarged to include all 3 million forenames and surnames. When we type a proper name, which the system will recognize as it begins with a capital letter, it allows say five characters to be keyed before it offers a pop up list of TTT names that begin with those five letters. The user compares the names offered with that required and selects the correct name module from the list. For names not in the list the user may complete the entry by keying, which the system will capture for incorporation into the published list.

There might be some talk of cost and difficulty. The spellchecker register has to be expanded substantially but the technology and cost-effective storage is available. With this in place there will be no excuse for not spelling someone's name correctly. However, personal letters are

small beer compared with the real benefit of ensuring that major commercial databases are error free. There are many ways of getting data into computer systems but keying is still the principal means for original capture of names.

Keying technology has changed somewhat and the load has been reduced by automatic capture of transactions, but corporations still have a number of keyers. Keying names is not always easy, even common names such as *Smith* and *Kenneth* are difficult because of the final *-th*. Keying data is often an entry-level job that is assumed not to require language skills mainly because of the previous emphasis on numbers: item numbers, item cost, numbers sold rather than names like *Ponce-de-Leon*. Management has placed the integrity of its personnel, customer and other databases into these hands, and has no easy way of checking that integrity. Examination of the names in such databases shows manifest errors. These errors cost money, and as we have seen, proper names are not in the spellchecker's repertoire so we cannot use the after-the-event process. Even if we could, how would we know which was right? We have to have control at the point of entry where we can compare what is being captured with the original document. Of course that is part of the current process but the new process, with its change of emphasis from creation to comparison, offers significant improvement.

How can a register be developed and used to improve the efficacy of name capture and replication? I don't know, although I have some ideas. What I do know is this is meat and drink to the American Name Society and the Canadian Society for the Study of Names. I suggest that the societies set themselves a challenge, marking the fiftieth anniversary of the founding of ANS, to create in the next 5 years "An American Register Of Names" (AARON) for use throughout industry, government, academe and genealogical study in North America and beyond.