The human use of human indexers*

Laurence Urdang

Every time a new invention becomes available, people worry lest they lose their employment. Undoubtedly, the invention of the wheel threw sledge-makers out of work (though we have no record of any singularly Luddite activity that far back); spinning jennies idled spinning wheel manufacturers (and users); the advent of the internal combustion engine destroyed the business of buggy-whip makers, just as the spread of railways reduced the number of stirrup cups and yards of ale served up to coachmen; Alexander Graham Bell may well be said to have been responsible for the extinction of the carrier pigeon. To bring the problem closer to home, the introduction of automatic typesetting has occasioned newspaper compositors' strikes all over America. In the long run, virtually all inventions have worked to the good of man, though those involved in the short run have undoubtedly suffered.

Certainly, the most significant technological advance in the middle years of the 20th century has been the development of the digital computer. Yet, it has frightened bookkeepers and clerks in much the same way that the early motor cars frightened horses.

The purpose of this article is to demonstrate that if computers are properly applied to indexing, not only can threats of replacement be obviated but more and better indexing can result. The proper use of the computer (as with any machine) is to relieve people of onerous, time-consuming, boring tasks, that can be done better and with greater uniformity and consistency by machine. This should have two effects: first, to allow indexers to concentrate their efforts on those phases of their work that are best done by humans (because computers are incapable of value judgments); second, to increase the number of books to be indexed (and the quality of that indexing) at a far lower cost than publishers are now paying. The expression 'information explosion' became a cliché about fifteen years ago and is scarcely heard any more. Yet, we are today virtually inundated by the debris from that explosion: in fact, it wasn't an 'explosion' at all but an unstoppable chain reaction. Information as an end in itself is, of course, pointless; but useful information that is rendered inaccessible by the lack of proper indexing tools is utterly worthless.

Thus, indexing must expand in direct proportion to the increasing availability of information in order for the information to have any value at all. It may be taken as given that the number of qualified indexers at any particular time is finite. If that be the case, then the work load for the individual indexer should be increasing as well, and we must seek ways in which to relieve the burden imposed by the routine aspects of indexing. The computer, correctly applied, is the solution, for it can leave the indexer free to do what he does best—the qualitative evaluation and interpretation of information. Truly, the computer can lead the way to the human use of human indexers.

As all indexers are aware, their work generally falls into two categories: the interesting part and the drudgery. The drudgery—writing out slips, alphabetizing, etc.—needs no further description here. The interesting part—content or interpretive indexing—probably requires no explanation either, but since it is the meat of every indexing project that affords the opportunity for an indexer to exercise intelligence, understanding, sympathy, empathy, discretion, discernment and creativity, it would be useful to consider it as the really important contribution made by the Indexer as a Human Being.

First, though, I should like to consider the Indexer as a Machine, or, to put it in the preferred order, the Machine as indexer.

The companies with which I am associated have been involved in the preparation of reference books of all kinds. Although we research and compile general and specialized

* With apologies to Norbert Wiener.

The Indexer Vol. 11 No. 3 April 1979 125
WEATHER
aerographics the study of atmospheric conditions. Also called aerography.
—aerographer, n.
climatology the science that studies climate or climatic conditions.
—climatologist, n.—climatologic, climatological, adj.
nephology the scientific study of clouds.—nephologist, n.
telemeteorography the recording of meteorological conditions at a distance, as in the use of sensing devices at various points that transmit their data to a central office.—telemeteorographic, n.
See also air and climate.

WIFE
uxoricide 1. the murder of a wife by a husband.
2. the husband who murders his wife.—uxoricidal, adj.
viduage the condition of widowhood. Also called viduity.—Obsolete, vidual, adj.

WINE
oenology, enology the science of making wines. Also called viniculture.
—oenologist, n.
oenomancy, oinomancy a form of divination involving observation of the colors and other features of wine.
oenophily, enophile the love of wine; connoisseurship concerning wines.
—oenophile, n.
oenophobia, phobia a dislike of or hatred for wine.—From Greek.

Figure 1. -Ologies & -Isms is organized alphabetically under a number of thematic categories. © Copyright 1978 by Gale Research Company.

dictionaries that are self-indexed, we also work on reference books that require an index in order to access certain kinds of information that they contain. Generally, such information is available only to the user who takes the trouble to read through an entire work in order to familiarize himself with its content—a ridiculous task, especially if you consider that the functional category reference book means 'a book to which one refers for information'. Perhaps it will be useful to cite several examples of the type of book I am considering. All of the books cited either have already been published or will be published during the first months of 1979, lest you suspect them of being apocryphal figments of my imagination.

Computer manipulation of terms
-Ologies & -Isms, Gale Research Company, 1978. 277 pp. From the Editor's Foreword: "-Ologies & -Isms is a dictionary that, on four counts, deserves the qualifier unusual. Its lexicon is drawn mainly from words ending in one of four suffixes, its organizing principle is thematic as well as alphabetic, it concludes with a unique index, and it is intended primarily for those who seek assistance in using the resources of libraries...

So much for the rationale. The organization of the book can best be seen in the excerpt reproduced at Figure 1. In this book, which is fairly simple in contrast to many of the reference books we prepare, there are only five basic kinds of information: thematic category; headword; definition(s); run-in sub-entry showing a reflex of the headword with a common suffix; cross-reference(s). Because the headwords are not readily accessible unless one knows the thematic category under which they are likely to occur, an alphabetical index of all headwords solves the problem rather neatly. In this instance, as can be
Germanophobia, phobias
geroformy, OLD AGE
gerodontics, OLD AGE
gerodontist, OLD AGE
gerodontological, OLD AGE
gerontology, OLD AGE
gerontophile, sex
gerontophobia, sex
Ghibellinism, POLITICS
Gigantism, size
Gigmanism, attitudes
Glaciological, geology
Glaciologist, geology
Glaciology, geology
Glossographer, words
Glossography, words
Glossolalia, speech
Glossolalist, speech
Glossology, linguistics
Glossophobist, phonograph records
Gothicism, art
Gothicism, literature
Gothicism, architecture, art
Gothic, architecture
Gothic, art
Gothic Revivalism, architecture
Gourmandism, behavior
Gourmetism, attitudes
Gradualism, philosophy, politics
Gradualist, philosophy, politics
Grammar, magic
Grammar, grammar
Grammatism, grammar
Grammatology, words
Grammatist, grammar
"Figure 2. A sample of the index from -Ologies & -Isms. ©Copyright 1978 by Gale Research Company.

seen from the section of the index reproduced at Figure 2, the referent need be no more than the thematic category. Running heads provide a quick orientation for the user to the alphabetical interval of thematic category offered on each page.

Such an index requires no creativity or imagination once its format has been established. Because the main text of the book was keyboarded into machine-readable form for automatic typesetting, the index could be generated automatically by computer merely by designing a program that extracted all bold face headwords and run-in entries from the text, along with the (capitalized) thematic categories, allowing the computer to "attach" to each of the headwords and run-ins the thematic category word, and then alphabetizing the resulting list to produce the index. A simple subroutine of the program rested the index; another allowed for the inclusion of more than one thematic category under a listing when the listing appeared under more than one category in the text. Thus, for example, autecology, which appears under both biology and environment, appears in the index as 'autecology, biology, environment'.

As far as I can see, there is no advantage, either to a publisher, who is trying to keep his budgets within reason, or to an indexer, who has more productive ways to spend his time than by drearily writing and alphabetizing slips, in having such dog-work done by human beings. The result of the computer-generated index is not only achieved far more rapidly and accurately but at lower cost: even the costs of keyboarding the index are saved since it is generated from data already in the computer. Also, although the final result must be checked to make certain that the program has run properly, there is no need to proofread it closely, thus saving time and money. In this respect, it must be remembered that in most cases, by the time a publisher has brought a book to typesetting, his investment in the project is approaching its maximum, and the sooner the book can be published, the sooner he can begin to see a return on his money. For this reason, we find that more and more complicated
and expensive reference projects are being keyboarded into machine-readable form as the editorial work progresses, instead of waiting till it is complete.

**Whole-text indexing**

Before describing more complex books that require both computers and human intervention in order to provide them with indexes of some quality, I should like to cite another rather simple example but one that by its sheer volume can awe any indexer. We recently had occasion to prepare a typical (though small) thesaurus based on the structure devised by Roget. Everyone is familiar with the organization: approximately 1,000 sets (or paragraphs) of synonyms organized by theme, accompanied by an alphabetical index of all the words listed.

In this instance, our editors prepared the main section of the book and it was keyboarded into machine-readable form. The entire index—which, in the end, constituted almost exactly half the book—was generated automatically. Here, of course, the referent in each case was the number of the synonym set (as compared with the thematic category in the preceding example). The saving realized by generating this index was dramatic on two counts: first, because every word of the text had to be indexed; second, because in the end only half of the book had to be keyboarded. Add to this the consideration that the time required for extracting and sorting the index can be a matter of computer minutes (depending on how powerful a computer is used). Would anyone care to estimate the man-hours required to index a thesaurus of, say 250,000 words? Or the costs of keyboarding all of those words all over again? Or the costs of proofreading the result?

**Cumulative indexing**

Obviously, computers become more dramatically cost- and labour-efficient the greater the quantity of data to be processed, and it would seem foolhardy to tool up a high-powered device to process an index of only a few hundred entries. However, that is not always the case, as I think I can demonstrate with the following example. Because the project is still in work, I cannot reveal full details, but the principle should be clear. We shall shortly undertake the compilation and publication of a quarterly 'reference periodical', each issue of which will number about 192 pages of 8-point type in a 9in x 6in format—just to give you an idea of the amount of text. The content will be in alphabetical order within each issue. In order to provide ready access to the information, the second, third and fourth issues (completing the volume) will contain cumulative indexes. There will not be an unusually large number of alphabetized headwords in any one number of the periodical—perhaps not more than 2,000—but it becomes quite worth while to employ a computer to prepare the indexes when one comes to consider the integration of the 2,000 from the first number with the 2,000 of the second, the addition of the third and finally bringing the total to approximately 8,000 listings. Were such an index to be done manually, it might not be too formidable; but the costs either of resetting the expanded index each time or of interfileing the successive new listings in hot metal would be very high. The plan calls for integrating the indexes of Volumes I, II and so on, so one can easily see the advantages of computerizing the entire system.

**Human help required**

I should like to turn now to books of considerably greater complexity than those discussed—reference books whose indexing can be greatly enhanced in every way by the application of the computer but which cannot be indexed solely by such a device. First is a work to be published early in 1979 by Verbatim called *Word for Word*, by Edward C. Pinkerton (Figure 3). This book is a little difficult to describe. Essentially, it is a book 'about' etymology, divided into 60 chapters each of which deals with a set of English cognates. The idea is to describe, in running text, words that share a common root. One chapter, for instance, deals with words like shipwright, millwright, playwright (etc.), boulevard, bulwark, ergophobia, energy, organ, organism, orgy, write, liturgy, dramaturgy, George, argon, irk, demiurge, surgery, and work, all of which share in common the Indo-European root *werg-* in a variety of its reflexes.

The book examines these words from the standpoint of their phonological, morphological, and semantic development, and, as can be seen from the sample shown at Figure 3, packs an enormous amount of information into almost every line of text. In fact, the density of information is so high that it was decided to number the lines of the book for detailed reference. In order to simplify indexing, only one numerical sequence is to be used, despite the fact that this would result in some five-digit line numbers.

The Indexer Vol. 11 No. 3 April 1979
The word thaumaturge was formerly applied to saints who possessed miraculous powers; it is now used more often to refer to magicians, conjurers and illusionists. In the same group we have liturgy ‘rites prescribed for public worship; formal religious worship conducted by authorized officiants for the benefit of the laity’ [via Late Latin liturgia from Greek leitourgia ‘public service; divine service’ from leitou ‘belonging to the public, lay’ from leis, laos ‘people’ + ergon ‘work.’ If a dramaturge writes plays, a metallurgist works with metals, and a thaumaturge works miracles, what does a demiurge do? The word démiourgos in ancient Greece meant ‘(one) working for the common weal; a craftsman, workman, artist; maker, creator’ [Greek démos ‘public’ from démos ‘common people’ + ergon ‘work’]. Plato used Démiourgos (capitalized) to mean ‘the Creator of the material world’ [Late Latin démiūrghus > English demiurge ‘a public official in certain ancient Greek states; the creator of the material world (Gnosticism).’ In the Platonic system the material, sensible world is vastly inferior to the ideal world of transcendental universals, and so is the Demiurge inferior to the Supreme Being.

The Greek element ergon ‘work’ also lurks in the English word surgery etymologically ‘manual labor; hand-work; the medical treatment of injury, deformity, or disease by manual and instrumental operations’ [from Medieval French surgerie, serurg(er)ie, from Old French chirurg(er)ie; from Latin chirurgia, from Greek kheirourgia ‘a working with the hands’]. Compare kheir ‘hand,’ as in chiropractic.

...
a. "weary, tired; bored"
b. "source of annoyance"
c. "latent power"
d. "exercise of power, actual working"
e. "public"
f. "cultivated lands"
g. "an inert gas constituting approximately one per cent of the earth's atmosphere"
h. "wild party"
i. "to swell"

There are others, I am sure.

Items like 'public' present no problem; but items like b, c, d, g, certainly require work in order to put them into acceptable indexable form. The computer can extract these quickly and painlessly, but it is incapable of making qualitative decisions or, for that matter, linguistically sophisticated decisions of any kind at all. The computer cannot do what the skilled indexer finds it second nature to do: content indexing. Therefore, either at the same time as the line numbers are being inserted into the printout of the extracted raw index or in a separate pass (by another indexer) through the raw data, the indexer must introduce such editorial changes as the decision to break item a. (above) into three listings:

'weary'
'tired'
'bored',

to change item b. into
'annoyance, source of,'
and item i. (which would otherwise end up under the T's when the alphabetization program is run) into a styled format—one that I prefer—that retains the order 'to swell' but ensures its listing in the S's.

I should now like to examine another sort of work that presents indexing opportunities somewhat different from those accorded by the preceding example. In 1979, Gale Research Company will be publishing a Dictionary of common suffixes in English (to be followed by a Dictionary of technical suffixes in English), prepared by Laurence Urdang Inc. On the face of it, this may not appear to be a particularly unusual work—till one looks at the arrangement of the information it contains. Because we felt that users may not always know where a given suffix begins—Is the suffix -logy or -ology? -a or -ia?—it was decided to list the suffixes in order, alphabetized from the end of the element, that is, from the right. Thus, a sequence like

-ly
-phia
-phobia

appearing in that order if alphabetized normally, would appear as

-phia
-phobia

-ly

when alphabetized from the right. Although we were captivated with the logic of this decision, we realized at the same time that people can find 'reverse alphabetization' somewhat confusing (and even infuriating) unless they are accustomed to it, and, to the best of our knowledge, the opportunities for acclimatization are few. Therefore, an index showing the suffixes in normal alphabetical order was required. But what would the index contain? '-ia see -ia' or '-philia see between -phobia and -dy'? Scarcely. It was obvious that the solution was to number the entries consecutively, and the referent then became '-ia 14', or whatever it was. The intervals between numbered entries were given at the fore-edge of each page in the running head. A further step had to be taken because some of the suffixes have more than one gloss. For example, the entry for -ia1:

14-ia1 A suffix borrowed from Latin with specialized meanings in medicine, zoology, and botany and a single general meaning.

1. Used in medicine to describe specific diseases, either physical or mental: pneumonia, hysteria, phobia.
2. Used in botany to denote a specific plant genus: Wisteria, Buddleia, Fuchsia.
3. Used generally to denote things derived from, belonging to, or related to a specified thing or person: Marylandia, tabloidia, Einsteinia.
4. Used in the names of taxonomic divisions to indicate classes or orders of plant or animals: Amphibia, Mammalia, Cryptogamia.

If -ia1 ends up with the number 14, then references to its subentries would be listed as 14.1, 14.2, etc.
Now let us take a look at the indexable information within an entry, using -ia1 as a fairly typical example. First, there is ia1 itself. Second, there are the sample words pneumonia, hysteria, phobia, etc. The computer program should encounter no difficulty in extracting and alphabetizing those. But the core of the definitions, which we should also want to access through the index, are another matter. For definition 1. we should like to see something like ‘disease, specific, 14.1,’ ‘disease, mental, 14.1,’ and ‘disease, physical, 14.1’ all listed in the index in proper order. The computer cannot accomplish such content or interpretive indexing. In addition, there are glosses, similar to those required by the etymological work, appearing in inverted commas, and these must be subjected to the same scrutiny and evaluative classification.

**Division of labour**

I have attempted to demonstrate that an enormous amount of the drudgery of indexing can be assumed by the computer—simple, straightforward extraction, limited editing (straight inversions, for example), alphabetization, and a few other routine processes that bore indexers to tears and fill them with dismay. On the other hand, there are certain functions that the computer is incapable of performing, and it is those that the indexer must continue to do. The point is that the computer should be used to do the mindless work that it is best suited for; it may upset some indexers that I am suggesting that they be replaced by computers, but I must confess that if the sort of indexing such indexers have been doing can be done more quickly, cheaply, and more accurately by computer, then so be it. However, I believe such instances to be of increasing rarity: the effect over the longer term will be to improve indexes of all kinds. After all, the purpose of an index is to provide the easiest access to the information contained in a book. In fact, it might be said that if a book containing information has no index (or other arrangement) or an index that, in effect, blocks the access to the information, then what is the point of publishing at all?

In conclusion, I should like to look at a kind of computer-oriented indexing that combines most effectively all of the virtues of all systems. In some kinds of books that are going to be keyboarded for automatic typesetting, it is sometimes possible to introduce the index coding either in manuscript or after the first text printout has been completed. In such cases, the indexer marks key words and phrases with a symbol that is keyboarded along with the text or inserted afterwards. (If the text is highly styled, with a great variety of boldface, italic, and other devices to classify information, then we are talking about one of the examples already discussed above.) This procedure is useful in encyclopedias, textbooks, and other non-reference works that are not quite so rigorously styled. The code inserted by the indexer is keyboarded with a symbol that never appears in the final typeset version because a computer program subroutine suppresses it. However, it is still there for the computer to ‘use’ for processing. If paraphrasic (content) indexing is required, it can be inserted in the margins and keyboarded, accompanied by the same kind of bracketing code that will ensure its not passing through as typeset matter. The computer can then extract the index matter. The next steps are to insert page (or other) references into the computer printout (after the text has been set), alphabetize the result, format it for setting and, finally, run the index through the typesetter.

The systems and procedures described here are intended to provide better indexes, with the various kinds of information they contain prepared by the most efficient, inexpensive, and expedient means. Computers should be relegated to doing the things they do best and most economically; indexers should be free to concentrate on those aspects of their art in which human beings are unequalled, not only by machines but by all those who turn to specialists when they want to ensure a job’s getting done right.

Laurence Urdang’s passport identifies him as a lexicographer. His linguistic and lexicographic pursuits have led him to the founding and development of Laurence Urdang Inc. (Essex, Connecticut) and Laurence Urdang Associates, Ltd. (Aylesbury, England), publishing companies devoted to the creation, compilation, design, computerization, and composition of dictionaries, encyclopedias, almanacs, and other kinds of reference books. In addition he is the editor of Verbatim, a quarterly for non-specialists about the English language. Mr Urdang divides his time between England and America.