Doing Other Things with Texts
The Use of Electronic Resources in Revising the *OED*

Jeffery A. Triggs
Oxford University Press

1993

There is much about electronic text and dictionaries that is still paradoxical.[1] At the time of the publication of *OED2*, one might fairly have said that it was a project with one foot in the nineteenth century and the other in the twenty-first. That the *OED* had begun conversion to electronic form — and SGML no less — as early as 1984 was nothing short of visionary. And a good portion of the work on the *OED2* text had been done on IBM computers. But the daily work of the lexicographers in Oxford went on much as it had in Murray's day, or indeed in Dr. Johnson's. The old building in St. Giles was still visited regularly by readers bearing bundles of 4 by 6 index cards, which were alphabetized by catchword and filed by hand. Ranges of these files were regularly "sorted" by hand. The sorters worked at computerless desktops with a volume of the dictionary open on a wooden book-holder, a wooden box of slips, perhaps a foot and a half in length, off to one side, and a mess of slips spread out puzzle-wise in front in the rough form of an entry. In this way, it was determined which quotes represented new words or senses, which were already covered by *OED*, and which would go to keep so much of James Joyce company in the one-offs file. The "new word" quotes would then be fastened with paper clips and put forward for drafting. The Oxford editors also worked with 4 by 6 cards, drafting their definitions by hand and bundling them with the supporting word-slips. Typesetting instructions — a single underline for

 This essay was first given as a paper at the Conference on Early English Databases, University of Toronto, Toronto, ONT, October 9, 1993.

italics, double underline for smallcaps, a squiggly line for bold — were added in pen to these same slips. If library research was necessary, the 4 by 6 cards would be mailed to libraries in various parts of the world and returned by mail suitably annotated. What the printer (eventually the keyer) got was literally held together by paper clips, or, in the case of larger entries, rubber bands. The greatest concession to technology seemed to be the allowing of whiteout, which as late as Burchfield's tenure had been banned in favor of hand-copying in the event of a mistake. By 1989 there were a few computers in evidence — an old IBM in the basement used to run occasional and expensive searches of Nexis, and a few Suns on which the new *OED* could be *looked at* by editors of the *SOED*, who nonetheless drafted their entries by hand.

I have gone into all this detail, not so much to poke fun at the old ways, which had the undeniable virtue of *working*, but to give some idea of how far we have come and a better sense of how far we intend to go as we push forward toward a completely revised *OED3*. It is worth remembering that *Webster's Third New* International has only quite recently been converted to electronic form (for us a most hopeful sign). The old ways of lexicographers were good ways, and should not be given up lightly. And yet, if we are to accomplish the enormous task of a full revision of the *OED* in less than 20 years, new ways of integrating computers into the editorial process will have to be investigated and put into practice.

Since the publication of *OED2*, there have been a number of significant advances in the use of computational tools in Oxford. For one thing, the e-texts of *OED* and the other Oxford dictionaries are now generally available to editors on a network of Sun workstations. A number of projects, beginning with the recent editions of the *ALD* and *COD*, have been edited completely on computer, and such editing is planned for *OED3* as well. A comprehensive, interactive editor is being tested even now. Both the English and North American reading programs are now fully electronic. Augmenting the old paper word slips accessible only by "catchword", we now have well over 600,000 electronic "slips" in SGML form that can be searched in seconds for any word in the text, in effect a corpus of over 15 million words drawn from several thousand discrete sources. The advantages of searching full texts should not be underestimated. As early as the appearance of the first trial CD-ROM of the *OED*, it became apparent that full text searches of the quotes already in *OED1* would yield many internal antedatings and other hitherto "hidden" items of interest. For example, the earliest quote for the "new" intransitive sense of "foreclose" was found in the entry for "mortgagee". Many times, the quote used in a new entry was taken for some other catchword and would thus have been lost to the old methods. Full-text electronic searches have

also enabled us to begin systematically checking and if necessary revising Murray's variant form paragraphs.

The incomings corpus, which of course has been skewed heavily toward contemporary sources, has already brought in dividends in the area of "new words". Recently, in anticipation of the revision stage and to add depth to the new word resources, both reading programs have extended their coverage to historical sources, especially the non-fiction literature of the late nineteenth and early twentieth centuries. In addition to this, we have explored other electronic resources that might be termed "historical".

Among these are electronic library catalogues, like the University of California's MELVYL, which contains, along with much else, a sizable corpus of historic electronic text in the form of titles. These may serve as quotations themselves, or point us toward textual quotations. For example, the title of a MELVYL dissertation provided us with the first quote for computational "parallelism"; another dissertation title led us to the stacks of the Harvard library for the first example of "reflexivity"; a number of titles found in MELVYL led us to sources containing antedatings of the attributive use of "broadleaf". To speed up the library research process, we've begun using email whenever possible.

The revision has also been aided greatly by the acquisition of the Michigan Early Modern English Materials in electronic form. These amount to over 50,000 electronic citations: old *OED* slips and slips collected by Fries which were keyed in as far back as the 1960s thanks to the foresight of Richard Bailey. They are now in use at Oxford, held as a special corpus and searchable in the same way as the incomings corpus. It is hoped that in the near future we will be able to make similar use of electronic resources from the *Old English Dictionary* and other historical projects, which should prove in unexpected ways the wisdom of Craigie's long ago having cast his bread upon the waters.

On another front altogether, we have forged an alliance with Lou Burnard of the Oxford Text Archive and have begun amassing a full-text historical corpus in TEI-conformant SGML. In only the first year, the historical corpus has grown to over 12 million words ranging in time from the early English alliterative poems through the early twentieth century. These are held together for searching as a special corpus in Oxford, but available as individual texts to scholars in general through the OTA. The historical corpus has a number of uses. It has already provided some remarkable antedatings, particularly of now recognizable new senses of core words such as "neglect" and "dearly". It is also a resource for revising the balance of *OED* citations — in favor, say, of greater American representation or greater representation of women authors. It has also served as a means of check-

ing quotations already in the *OED*. Folio and quarto texts of Hamlet, for instance, enable us to tell automatically which text has been the source of an *OED* quotation and to address the dating of these quotations with greater specificity.

Beyond accumulating electronic resources, however, dictionaries need to consider how these resources can be used. String searching is the most obvious technique. It can take the place of tedious combing of paper files and concordances, and indeed with greater efficiency. But unless it is made part of an interactive editing system, the paper process is essentially unchanged. We need to find equivalencies, not merely for the tedious parts of paper lexicography, but for those accomplished comfortably with paper. Paper word slips had the great benefit of homogenizing texts of all sorts into a form easily manageable by an editor: quotations could be included or deleted simply by shifting from one paper clip to another; they could be sent out for research simply by popping them in an envelop; the bundles packed into a long wooden box contained full text and instructions for a printer to make up a page. Anyone who has experienced the mess of texts available in machine-readable forms (and I use the plural advisedly and sadly) can appreciate the importance we place on homogenizing these texts and the value of the TEI's scheme of encoding. An editor must be able to incorporate text automatically, and thus be able to manipulate text at will. This implies an easy interaction with the operating system and an internet as opposed to a "closed" CD-ROM environment for anything but supplemental uses. The dictionary office of the future may involve editors working at many different sites accessing their materials through remote connections and communicating through email. UNIX has the greatest potential for providing such a flexible environment.

A big advance would be the development of electronic "sorting" in the lexicographical sense — that is, a means of filtering huge full-text corpora to flag potentially interesting items without human intervention at an early stage. A word-frequency cut-off, like that used in *COBUILD*, would not do for the *OED*. The *OED* user is not presumed to be a "learner" interested primarily in serviceable contemporary vocabulary: "unhandselled" or "vastate" have their places in our word-list along with "supermarket", and obsolete senses of interest, say, especially to Shakespearean scholars cannot be neglected. One might easily generate a list of strings that do not appear in the text of the present *OED*, but such a filter, call it a spell-checking type, could not account for minor senses of polysemous words, nor could it address sufficiently the thorny issue (pun useful here) of variant spellings. A recent hypertext experiment with the Folio Shakespeare at Bellcore was frustrated when it tried to include an *OED* "button" fetching *OED* entries for a given word by headword — the practical solution being to change to

a modernized text. One promising but as yet vaporwarish approach involved using bigram or trigram word frequencies to generate a "real-word error" spell list. One problem is that most computational studies limit themselves to modern texts and are skewed toward linguistic rather than lexicographical uses. Broad-based theory is certainly helpful at the overall planning stages of a project like the *OED*, but the lexicographer is still confronted daily by the slow march of his "range", by the tyrannies of the alphabet and the budget, and the emphasis must be on what is intrinsically useful. To borrow a Rilkean formulation, practicality is everything.

Early English databases can be useful in a number of ways. Any full-texts or electronic dictionary materials, such as MEMEM, can be incorporated directly into the working environment by means of specialized search engines like Open Text's Pat and the basic UNIX utilities. Electronic glossaries and word lists present a special opportunity and a problem. It is our policy to avoid including words for which there is no contextual evidence *other* than inclusion in a glossary. There are quite a few words in *OED* and other dictionaries on the strength only of a single use by Shakespeare — indeed, an interesting paper could be written on the topic of the influence of dictionaryless Shakespeare on later dictionaries. A special case could no doubt be made for Shakespeare, but there are many other authors in this category as well. The *OED* entry for "adyt", a variant of "adytum", is based on a single quote from Greene and until recently when a use turned up in a book on Chesapeake Bay we had no other examples. Not long ago, the *OED* staff was flooded with requests from the 8th *COD*, which was re-examining some of the variant lemmas included in the original *COD* by Fowler. A number of these were not in *OED* and had been carried over from edition to edition without supporting textual evidence. To get a sense of the potential size of the problem, one need only consider the prospect of researching and drafting the exotic, one-off vocabulary of *Finnegan's Wake* on the strength of Joyce's mention alone. On the other hand, electronic glossaries and word lists, while not sufficient in themselves to justify a word's inclusion in *OED*, might act as filters for searching full text sources. Thus, they could provide a kind of immediately practicable electronic "sorting", adding an exciting new feature to the revision process. There are still problems here, but they are not at all insurmountable.

In conclusion, I would like to stress that dictionaries will not soon if ever be generated automatically — the human element cannot be dismissed. The goal should be to supplement or find equivalents for the old methods and to toss out only those elements of the process, such as snailmail communications or handwritten copy, that are clearly unsuitable in the modern age. What is needed is not so much fancy new hardware, or GUIs, or crazy hypertext links that depend

on a particular platform, but flexible software, freedom of access to standardized, transportable texts, and freedom of manipulation.