The *New OED* project at Waterloo:
old wine in new bottles

by D. W. Russell[1]
Waterloo Centre for the New OED
University of Waterloo

In mid-1983, three years before the fourth and final volume of the *Supplement to the Oxford English Dictionary* (OED) appeared in conventional print form, Oxford University Press (OUP) had announced its intention of computerizing the *OED*. By the end of 1983 the decision had been made that Oxford University Press would manage this large undertaking itself, and would enter into a series of contracts/agreements with other parties (such as the University of Waterloo, IBM UK, International Computaprint Corporation (ICC) of Fort Washington, Pennsylvania, among others) to carry out work on various aspects of the computerization.

From a long-term perspective, the Project's objective is to transform both the original 12 volumes of the *OED* and the 4-volume *Supplement* into an electronic database; the work to be done in arriving at this objective has been broken down into several discrete phases: first, the initial sixteen volumes had to be entered into the computer, preserving the original text organization and presentation. This meant the tagging of structural and typographical elements in the dictionary as the data were keyed and transfered onto magnetic tape. One of the aims of this phase is to produce a printed version of the dictionary, integrating the *Supplement* material with the body of the *OED*, and incorporating about 4000 new entries into this merged edition. This version will be printed in the spring of 1989, in a projected 22 volumes at a price of about £1,500. The next major phase involves the design of a database structure for the machine-readable data, so that alternative structures can be presented, and interactive querying will be possible. It is at this point that the *New OED* comes into being, leading to expanded, updated, and revised versions of the dictionary that will allow several modes of data access, ranging from direct, online access, to the conventional printed version, and special, printed subsets of the dictionary. As can be seen, the possibilities are many, and the prospects for

[1]Presented at the International Association for Social Science Information Service and Technology (IASSIST) Conference held in Vancouver, British Columbia, Canada on May 19-22, 1987

lexicographical work seem to expand almost infinitely.

The magnitude of the proposed task can be appreciated when one considers the size of the current source data, the sixteen volumes of *OED* and *Supplement*: there are over 21,000 pages of three-column print, with a total of about 500 million characters including punctuation and spacing. This breaks down to about 306,000 main entries, 163,000 subordinate entries, with over 2,350,000 illustrative quotations, and almost half a million cross references. In addition to the sheer size of the dictionary, the material has, as anyone who has ever used the *OED* knows, an extremely complex structure, by which explicit and implicit information is conveyed to the reader through both the layout and the typography of the text. At the simplest level, each entry usually includes a headword lemma, a pronunciation key, a grammatical category label, a list of variant forms given by century, an etymological section, a series of sense definitions, each with a quotation bank; there is usually one quotation per century, with date, author, source and bibliographic reference. This structure is made more complex by words or forms that do not fit easily into the usual categories, and by the inconsistencies in method inevitable in a project that spanned many years and several generations of lexicographers.

Still, the capture of the dictionary material, which had to be done by manual keying of the text, since it was too complex for optical scanners, has been done, and done surprisingly successfully over a period of 18 months, ending in June 1986. The error rate, as noted by human proof readers hired by Oxford, is seven or fewer errors per 10,000 keystrokes. Working from enlarged copies of *OED* text, the key entry operators entered tags to identify all the typographical elements, as well as some of the structural elements of the source text. One of the chief aims at this stage, was, after all, to be able to produce a new, typeset edition in 1989.

But, since even this seemingly routine task becomes more complex when one is dealing with as much material as is found in the *OED* and *Supplement*, a further refinement was introduced at this point in the process. Computer science researchers at Waterloo developed a parser which automatically tagged structural elements not tagged during keying, and which converted the ICC tags into SGML codes.[2] This parsing permitted the automatic validation of the earlier tagging. The conversion to SGML codes was needed to permit a greater degree of automatic integration by computer of the *OED* and *Supplement*, as well as to facilitate the lexicographical team's interactive integration of the two source texts.

Meanwhile, at the University of Waterloo, research on database design for the *New OED* has begun, with the support of a 1.3 million dollar grant from NSERC. The Project has the mandate to design a database for the *New OED*, and to develop software utilities for database access, maintenance and update. This work is to be carried out over three years by a team of seven full-time researchers, directed by a team of computer science faculty researchers, led by Gaston Gonnet and Frank Tompa. To date, two prototype software tools have been produced, to allow interactive access to, and sophisticated querying of, the database. These tools are named *Goedel* and *Pat*. Although it is not within my competence to describe any of the technical specifics of these tools, I would like to offer some examples of the results currently possible, drawing mainly from my own research interests in the *OED*, namely the identification and study of the Anglo-Norman elements which were adopted by the English language in the medieval period.

[2]See Rick Kazman, *Structuring the text of the Oxford English Dictionary through finite state transduction*, (M. Math. thesis) University of Waterloo, 1985.

When the first tapes of the *OED* became available at Waterloo in 1985, it was theoretically possible to begin searching the dictionary interactively. In order to find all words labelled as Anglo–French or Law French, I needed simply to ask the computer to list occurrences of the relevant strings, "AF", "ONF", "Law Fr.", etc. At the time the data were mounted in a series of files, which meant the queries had to be repeated for each file. The results were not easy to read, and although one could scroll back or forward to identify the headword, in the end it proved easier to use the printed dictionary to locate the relevant entry.

After the creation of *Goedel* in 1986, a whole new range of possibilities made my querying of the data easier. I could now ask for the extraction of material according to structural categories in the dictionary entries, and according to specific strings within each category. I decided to limit my extraction to listing the headword lemma, the material within the etymological section, and the dates of the illustrative quotations, for all entries which could be considered to be derived from Anglo–Norman sources. Within the structure of the *OED*, Anglo–Norman material is identified in various fashions in the etymology section: it may be labelled as Anglo–French, it may be labelled as Old French (or Old Norman French or Old Law French), or it may be found to be labelled as French but with quotation dates preceding 1500. The results of my queries using *Goedel* could be printed in a formatted form which is eminently readable, and could be carried away for use elsewhere, freeing the researcher from being tied to a computer terminal. With the extraction power and flexibility of *Goedel*, the humanist researcher is faced with the new challenge of creating more sophisticated queries, based on previously unexamined possibilities, and building on the results of his or her ongoing interactive research.

The second extraction tool, *Pat*, allows a rapid and effective querying of a different sort, based in part on pattern matching; with *Pat*, for example, I could extract all entries derived from AF or OF and which have supporting quotations from a particular author, such as Chaucer or Gower. Or, to give another example, a researcher looking for infantine language was able to search the dictionary for all words whose sense definition included specific key words, such as "little" followed by "boy" or "girl" or "child" within a specified number of characters. Increasing familiarity with the results of these searches led to more sophisticated querying of the database, and raised technical problems which were addressed by the data structuring group working on the Project. In a similar way, *Pat* was used by a researcher interested in the source and frequency of quotations. It is possible to extract all quotations from a particular author, such as Fanny Burney, and further, to extract only quotations from Burney from a specified work, such as *Cecilia*.

These short examples, from among the preliminary group of research projects underway at the Waterloo Centre, serve to emphasize the futuristic nature of the Project; it is difficult to design a database for uses which may arise in the future, but which have not yet been imagined by humanist researchers. In an attempt to come to grips with this problem, OUP and the University of Waterloo conducted a user survey, to find out how individuals use the *OED*, to determine the principal facilities needed for the *New OED*, and to provoke considered responses about applications for the electronic version of the dictionary. Over 1,100 individuals were contacted, of whom 60% were from the UK and Europe, 40% from North America. The sample included both academic and non–academic users, with an emphasis on sophisticated users. The results have not yet been completely analyzed, but preliminary results give a broad outline of what users would like to see in the *New OED*. Basically, most

users want everything currently in the *OED*, in a fashion that is simple to use, quick to give results, and cheap to access. Ideally the data will be available publicly through an online data service, and privately via some disk format. The software must be user friendly, allowing quick access, both to a skeleton summary of each entry, and to a complete entry or selected details of an entry. The electronic version must not be prohibitively expensive, and yet be amenable to continuous updating and revision. And finally, the *New OED* should continue to be published in printed form. The survey results also suggest future applications, many of which will require revision of the *OED*: these include semantic field searches, frequency ratings by date or language of origin, the use of the *OED* as a thesaurus, and so on. One of the suggested applications, requiring the ability to search phonetics elements, will be made much easier by the decision to replace Murray's phonetic transcriptions with transcriptions based on the symbols used by the International Phonetic Association. This revision is to be incorporated in the 1989 print version of the *New OED*. Searching on these phonetic elements will only be possible, of course, in the electronic version of the *New OED*.

Just what the electronic form of the *New OED* will be has not yet been decided. There are plans to market an exploratory electronic issue of the old *OED*, without the *Supplement*, and without revisions, on CD-ROM in late 1987, with the aim of validating some of the results of the user survey, and generating feed-back for the database design currently in progress for the *New OED*. The database created to produce the print version of the *New OED* in 1989 is not now in a form which OUP would offer to outside users, but the Project does expect to market an electronic version of the 1989 and later editions of the *New OED*. The undertaking does present a number of technical and legal problems, the solutions to which will determine the final end product. And the planned revision and enhancement of the *New OED* after 1989 will certainly carry the Project well into the twenty-first century. Judging from the unforeseen shifts and changes in plan which plagued James Murray and the other editors involved in the creation of the *OED* from 1879 to 1933, it is perhaps unwise to promise completion of the *New OED* Project by any given date. But there is room for cautious optimism. Such is the position taken by Tim Benbow, OUP's Director of the *New OED* Project, in his Status Report given at our second annual conference at Waterloo in November 1986. He said:[3]

> "At the moment the project is running on schedule and on budget. We are not, however, complacent. Recurrent nightmares — one, in which as a juggler one is constrained to keep an increasing number of dictionary volumes of monstrous size in the air under pain of lexicution, alternating with a Sisyphean vision of ordering acres of dictionary slips only to have them taken by the wind as the last is about to be positioned — see to that!"

It is, no doubt, significant that Tim Benbow's nightmares do not yet reflect the presence of any textual database monsters.□

[3]Tim Benbow, "Status Report on the *New OED* Project." Paper given at "Advances in Lexicology", Second Annual Conference of the UW Centre for the *New Oxford English Dictionary*, Waterloo, Nov. 9-11, 1986.