[*Translating and the Computer 9*. Proceedings of a conference… 12-13 November 1987, ed. Catriona Picken
(London: Aslib, 1988)]

The *New Oxford English Dictionary* project

Timothy Benbow

Director, New Oxford English Dictionary, Oxford University Press, Oxford, UK

THE OXFORD ENGLISH DICTIONARY

The *Oxford English Dictionary* is the largest and most authoritative dictionary of the English language. It is a dictionary based on historical principles: that is, it takes as its subject-matter the entire vocabulary of the English language since 1150 AD. The *OED*, which is in twelve volumes, took approximately fifty years to prepare and the completed work was published in 1928. A *Supplement* to the Dictionary, on which work started in the 1950s, was published in four volumes between 1972 and 1986.

Almost half a million words are defined in the *OED* and its *Supplement* and the definitions are illustrated by over two million quotations. The vast size of the work, as we shall see, has an important influence on the way in which the *New OED* project has to be handled (see Table 1).

THE OBJECTIVES OF THE PROJECT

A work of reference like the *OED* requires continuous updating and revision to keep up with constant linguistic, social and technological changes. The publication of further supplements would be an inadequate, impractical and uneconomic solution to this problem. So would traditional paper-based methods of revision. Computerisation offers the only practicable solution. It also offers additional benefits in the form of a new and powerful research tool for literary and linguistic scholars and for other professionals in disciplines such as law and medicine, and for scientists, authors, translators and journalists: a lexical database of English.

The overall objective of the project is to create a machine-readable version of the *OED* and *Supplement* from which new editions of the Dictionary can be produced, in both printed and electronic forms. However, a task of this magnitude can only effectively be done in phases. First, we plan to publish a printed edition in which the entries of the *Supplement* have all been inserted into their correct places within the text of the *OED.* The aim of subsequent phases will be the setting-up, expansion, updating and publication of a *New OED* electronic database.

ORGANISATION OF THE PROJECT

The project is being run by Oxford University Press but at least three organisations have an important role in its execution:

(1) International Computaprint Corporation (ICC), a US subsidiary of Reed International, carried out the conversion of the text of the *OED* and its *Supplement* to machine-readable form.

(2) IBM (United Kingdom) Limited have donated equipment, software and expertise to facilitate the completion of the first phase of the project – details of which are given below. The value to OUP of this donation is in the region of £1 million. Of equal importance to the financial value of this assistance is the very considerable expertise in developing complicated software systems that IBM brings to the project. IBM seconded three experts to the project to help with the design of the computer system for Phase 1.

(3) The design of the database for the electronic version of the *New OED* has been entrusted to Canada's foremost university in the field of applications software: the University of Waterloo. Although the role of the University of Waterloo relates principally to the second and subsequent phases of the project, it made an important contribution to Phase 1: by seconding a computer scientist to OUP to develop a structural parser, the function of which is explained below.

Assistance for the project was also obtained from the Department of Trade and Industry in the form of a grant towards the research and development costs. OUP received a total of £285,000 over the first three years of the project. Again, the fact of the grant – in that it testifies to the Government's interest in a project of such cultural and technological importance – is almost as important as the *amount* of the grant.

Phase 1

1. *Principal activities*

Phase 1, as I have already mentioned, will take us to the printing of an integrated version of the present edition of the *OED* and its *Supplement* from an electronically-stored text. The principal activities involved are:

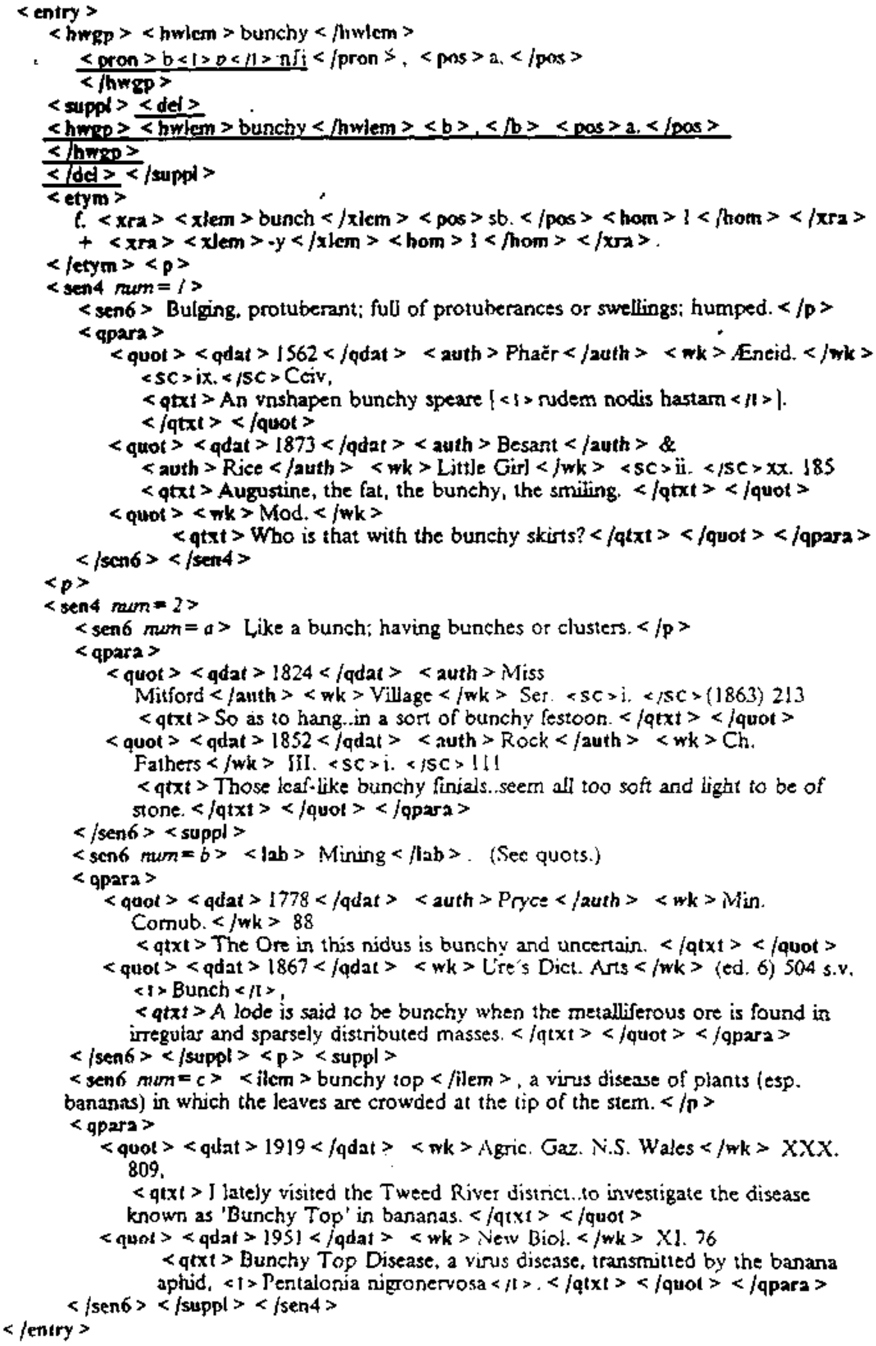
—structural analysis, to ensure that the logical structure of the very complex text is transferred successfully to the electronic version during keying
—keying of the entire text, together with additional structure codes
—checking, by using both computer software and conventional proof-reading, to ensure that the keying is accurate
—integrating the *Supplement* into the main body of the *OED*, initially by computer and subsequently, in those cases too complex for the computer, by online editorial intervention
—typesetting, by the transfer of the electronic text to a fast photo-typesetter
—printing and binding

2. *Structural analysis*

The text of the *OED* is both voluminous and complex. The statistics given in Table 1 bear witness to the *extent* of the text. The *complexity* of the text cannot so readily be illustrated. The fact that entries can vary in length from less than five words to more than one hundred thousand is an indication of this complexity. (See Figure 1, showing part of the entry for ‘bunchy’ in which the structure is indicated by SGML (Standard Generalised Mark-up Language) tags.)

Volumes (OED 12 + Supplement 4)	16
Pages	21,000
Words	60,000,000
Characters	350,000,000
Keystrokes	500,000,000
Entries	320,000
Illustrative quotations	2,350,000

Table 1. OED - extent



```
< entry >
   < hwgp >  < hwlem > bunchy < /hwlem >
      < pron > b < i > ʌ < /i > nʃi < /pron > ,  < pos > a. < /pos >
      < /hwgp >
   < suppl >  < del >
   < hwgp >  < hwlem > bunchy < /hwlem >  < b > , < /b >   < pos > a. < /pos >
   < /hwgp >
   < /del >  < /suppl >
   < etym >
     f.  < xra >  < xlem > bunch < /xlem >  < pos > sb. < /pos >  < hom > 1 < /hom >  < /xra >
      +  < xra >  < xlem > -y < /xlem >  < hom > 1 < /hom >  < /xra > .
   < /etym >  < p >
   < sen4  num = 1 >
      < sen6 >  Bulging, protuberant; full of protuberances or swellings; humped. < /p >
      < qpara >
         < quot >  < qdat > 1562 < /qdat >   < auth > Phaër < /auth >   < wk > Æneid. < /wk >
            < sc > ix. < /sc > Cciv,
            < qtxt > An vnshapen bunchy speare [ < i > rudem nodis hastam < /i > ].
            < /qtxt >  < /quot >
         < quot >  < qdat > 1873 < /qdat >  < auth > Besant < /auth >  &
            < auth > Rice < /auth >   < wk > Little Girl < /wk >   < sc > ii. < /sc > xx. 185
            < qtxt > Augustine, the fat, the bunchy, the smiling.  < /qtxt >  < /quot >
         < quot >  < wk > Mod. < /wk >
               < qtxt > Who is that with the bunchy skirts? < /qtxt >  < /quot >  < /qpara >
      < /sen6 >  < /sen4 >
   < p >
   < sen4  num = 2 >
      < sen6  num = a >  Like a bunch; having bunches or clusters. < /p >
      < qpara >
         < quot >  < qdat > 1824 < /qdat >   < auth > Miss
            Mitford < /auth >  < wk > Village < /wk >  Ser.  < sc > i.  < /sc > (1863) 213
            < qtxt > So as to hang..in a sort of bunchy festoon. < /qtxt >  < /quot >
         < quot >  < qdat > 1852 < /qdat >   < auth > Rock < /auth >   < wk > Ch.
            Fathers < /wk >  III.  < sc > i.  < /sc > 111
            < qtxt > Those leaf-like bunchy finials..seem all too soft and light to be of
            stone. < /qtxt >  < /quot >  < /qpara >
      < /sen6 >  < suppl >
      < sen6  num = b >   < lab >  Mining < /lab > .   (See quots.)
      < qpara >
         < quot >  < qdat > 1778 < /qdat >   < auth > Pryce < /auth >   < wk > Min.
            Cornub. < /wk >  88
            < qtxt > The Ore in this nidus is bunchy and uncertain.  < /qtxt >  < /quot >
         < quot >  < qdat > 1867 < /qdat >   < wk > Ure's Dict. Arts < /wk >  (ed. 6) 504 s.v.
            < i > Bunch < /i > ,
            < qtxt > A lode is said to be bunchy when the metalliferous ore is found in
            irregular and sparsely distributed masses. < /qtxt >  < /quot >  < /qpara >
      < /sen6 >  < /suppl >  < p >  < suppl >
      < sen6  num = c >   < ilem > bunchy top < /ilem > , a virus disease of plants (esp.
     bananas) in which the leaves are crowded at the tip of the stem. < /p >
      < qpara >
         < quot >  < qdat > 1919 < /qdat >   < wk > Agric. Gaz. N.S. Wales < /wk >  XXX.
            809,
            < qtxt > I lately visited the Tweed River district..to investigate the disease
            known as 'Bunchy Top' in bananas. < /qtxt >  < /quot >
         < quot >  < qdat > 1951 < /qdat >   < wk > New Biol. < /wk >  XI. 76
               < qtxt > Bunchy Top Disease, a virus disease, transmitted by the banana
               aphid,  < i > Pentalonia nigronervosa < /i > . < /qtxt >  < /quot >  < /qpara >
      < /sen6 >  < /suppl >  < /sen4 >
< /entry >
```

Figure 1. Extract from an entry — structure indicated by SGML tags

Considerable time had to be devoted to an analysis of the text to ensure that no information, either explicit or implicit, would be lost in the transfer of the text from the printed page into electronic form. From an early examination of the text it was clear that there was no simple and direct mapping from typography to structure: it would not be sufficient to code for typography alone since most typographical devices had differing structural duties depending on their position in the entry.

3. *Data capture*

Not only is the typography of the *OED* rather intricate, but the impression of the type itself is now blurred and broken. For these and other reasons optical character recognition was ruled out. Instead, the entire text of the *OED* and *Supplement* was keyed onto magnetic tape by keyboarders working from lightly marked-up enlargements of the source.

The keying of the text, which was carried out in the United States by International Computaprint Corporation (ICC), took eighteen months and was completed in June 1986. ICC sent us batches of proofs and magnetic tapes each month. The proofreading itself was a not inconsiderable task: the average monthly batch was equivalent in size to over forty 200-page academic monographs! The proofs were checked by freelance proofreaders under our control, marked up and returned to ICC for correction. The level of correction was expected to be low: ICC agreed to – and achieved – a target error rate of not more than seven keystrokes in 10,000. Data on the magnetic tapes was then transferred to our computer system for validation checking and processing.

4. *Computer system*

The project required a sophisticated electronic text processing system that would:

—accept and validate data captured by ICC
—hold the text safe, secure, and readily accessible
—provide a text editing facility for the correction, manipulation, and addition of text
—parse and add structural tags to the dictionary text
—integrate the texts of the *OED* and *Supplement*
—resolve cross-references
—convert the original phonetics to the International Phonetic Alphabet (IPA)
—format and output text on varying devices

Where possible, standard software products have been used. For example, we are using the IBM VM operating system and IBM's SQL database system to store the text. However, many parts of the system had to be developed by the

Project's computer group. The main tasks for which special applications programs were developed are:

—Parsing: It was not possible, during keying, to insert *all* the tags necessary to define the full structure of the text. A parser was therefore developed to tag those structural elements not tagged during keying. The parser had a second function: to transform the tags inserted during keying into SGML tags. This level and standard of coding was essential if a high degree of automation was to be achieved in the integration process.

—Integration: Those of the entries in the *Supplement* that are completely new have to be placed in their correct sequence among the *OED's* entries. The others, the incomplete entries, require fairly sophisticated handling. Each entry contains instructions about how the rest of the information within it should be grafted on to the equivalent *OED* entry; for example, *add, substitute for definition,* or *earlier examples.*

—Cross-referencing: With the amalgamation of the *OED* and *Supplement,* and the consequent shifting and altering of sections of text, many cross-references are rendered invalid. Whereas parsing and integration could be carried out piecemeal on small sections of the text, the cross-referencing problem could only be tackled once the whole of the text had been integrated.

—Pronunciation: The original phonetic notation devised by James Murray, which is now outmoded, has been converted to the International Phonetic Alphabet.

—Editing system: The complex nature of the text and of the processes, such as integration, to which it must be subjected in this phase of the project, requires that sophisticated online editing faculties be provided to assist the lexicographers in their tasks. The type of facilities required included: the use of colour to distinguish tags from text, the on-screen representation of a wide range of special sorts, and an integrity check of the text to ensure, before an amended entry is returned to the database, that the structural integrity of the entry has not been breached.

5. *New words*

Although the prime concern in Phase 1 is to integrate the *OED* and *Supplement,* we are also adding a limited number of entries for new words, and new senses of existing words, to the integrated edition.

Beyond Phase 1

In future phases of the project our principal aims will be:

—To publish an electronic version of the Dictionary.

—To update the Dictionary by adding in the hundreds of new words and meanings which each year enter the English vocabulary as a result of scientific, technological, social and political change.

—to revise the Dictionary, with particular attention to outdated scientific and technical definitions, uneven etymological information, subject classification of terms and other elements – yet to be determined – of a similar nature.

—To enhance the database by the addition of new and different material: national dictionaries (e.g. Australian, Canadian and South African); ELT (English Language Teaching) information; foreign language dictionaries; a thesaurus; illustrations; *OED* archive material.

1. *The electronic database*

An electronic database has been created for Phase 1 of the project. This database is a working tool for the production of an integrated edition of the Dictionary. It is not in a form that we would expect to offer to outside users, although we confidently expect that it will be possible to transfer the data from this working database to one tailor-made for the market.

The University of Waterloo, in Canada, has undertaken to design for us such a database and the necessary database management system to sustain it. As a first step, OUP and the University of Waterloo have jointly conducted a user survey to determine likely ways in which the database will be interrogated. The types of query that users of an electronic version of the *New OED* may wish – and should be able – to ask are illustrated in question 11 of the questionnaire (see Appendix). From this it is plain to see that a whole realm of information hitherto buried in the printed version of the *OED* will become readily accessible once the electronic version exists.

The response rate to the user survey was extremely good – over 50 per cent of those to whom questionnaires were sent completed and returned them. What was even more gratifying was that two-thirds of the respondents answered question 12, which asked them to give suggestions for further uses of an electronic *OED* not covered in the rest of the questionnaire. The University of Waterloo is currently analysing the information gathered by the survey and we hope to have the results very soon.

2. *Updating*

Undoubtedly the most vital and tangible aspect of updating is the addition of new words and senses. The Dictionary word-file was opened nearly thirty years ago, and it is now nearly twenty years since the most important new items in the early part of the alphabet were selected for the *Supplement.* The file is still growing, and there is now a large crop of recent acquisitions to the language (e.g. *database, day centre, deconstructionism, disinformation)* ready to be edited. A new words unit was set up in 1983 to deal with this entry preparation; its product is known as NEWS (New English Words Series). The unit currently has a staff of four. It provides continuity of method and of organisational infrastructure (specialist consultants, etc.) between the *Supplement* and the *New OED.* It is our intention to include a section of NEWS entries in the integrated edition now in preparation; thereafter, the material will provide a principal means of continually updating the *New OED* database. In addition, the entries will serve as a basic resource for other Oxford dictionaries in preparation.

3. *Revision*

Principal among the elements of the Dictionary requiring revision are:

Definitions. The definitions of current scientific terms, especially in chemistry, biology and physics, require revision in the light of present-day knowledge. This also applies to the treatment of technology and crafts, social sciences, psychology, economics, and the like.

There are many encyclopaedic references in the *OED* which are now inaccurate. The names of places and institutions have been changed, political systems replaced, and currency values rendered meaningless. Events now distant are described as recent.

Many definitions enshrine social attitudes which are discarded, especially racial, religious, sexual and class prejudices.

Etymology. *OED's* etymologies enshrine some old-fashioned principles and use outdated terminology: their transcription of alien linguistic forms is inconsistent. They contain a fair amount of incorrect information and lack the not inconsiderable findings of subsequent research.

4. *Enhancement of the database*

I have already mentioned some of the new and different materials that could be added to the *New OED* database: national dictionaries, ELT information, foreign language dictionaries, a thesaurus, illustrations and *OED* archive material. As with all the aspects of the project that are outside the scope of Phase 1, there are many questions yet to be resolved concerning the addition of non-OED material to the database. Two of the major ones are: what will be the structural relationship between the Dictionary and the new material (our current inclination is to view the separate entities as satellites to the Dictionary), and what pathways through the database should we provide?

5. *Potential products*

The range of potential products will be dramatically increased when the Dictionary exists in electronic form. The possible permutations are based on the three quite distinct elements which will go to make up each individual product:

(a) The *information:* the contents of the Dictionary or its satellites.
(b) The *medium:* the form in which the information is held (for example: in print, in an online database, on optical digital disks, on computer chips).

(c) The *software:* the programs which provide means of assembling, accessing and generating information.

Table 2 gives an indication of the range of products we are currently considering. It is merely indicative and we intend, as phase 1 progresses, to refine it considerably. A CD-ROM version of the twelve-volume *OED* (that is, the unintegrated text) is now in preparation and should be published by the end of 1987.

Plan:	FIRM	POSSIBLE	CONJECTURAL
Element: CONTENTS	Update OED Revise OED	Thesaurus National dictionaries OED Archive Material	Line drawings Foreign language dictionaries Moving images Audio material
MEDIA	Print On-line Optical digital disk	On-demand printing Microform Magnetic tape	ROM chip
SOFTWARE	Standard access Standard screen presentation	Subset generation Thesaurus access	Satellite data access

Table 2. Product matrix

6. *Potential markets*

As the range of potential products will have increased, so will the range of potential markets. The most basic distinction will be between the market for printed products and that for electronic ones. Current estimates indicate that the electronic sector constitutes at least 20 per cent of the total information market. This share seems set to grow at a rapid rate in the coming years.

There are currently a number of technical problems to be resolved. Among them are the problems of presentation (how will the complex text of the *OED* be displayed on a wide range of computer terminals?) and of compatibility (how can we ensure the widest compatibility of both data and software?). Given the rapid rate of technological advance, it may well be that these problems will be resolved by the industry before they become a pressing concern to us.

APPENDIX

User survey - question 11

11. An electronic version of the *OED* will have a wide range of applications. The following examples have been chosen to illustrate the way in which the basic dictionary elements might be combined in order to answer specific questions.

Please indicate those which would be of use to you. Do not hesitate to mark a question which, with appropriate changes, could be adapted to the discipline in which you are particularly interested. To give a couple of examples: question (d) could be adapted by *a* journalist to find examples of the suffix *-gate* used to name political scandals that resemble Watergate; question (h) could be adapted by a chemist to produce a similar list of chemical compounds.

Obviously, information yielded by many searches of this kind would not be comprehensive but could well form a valuable point of departure for further research.

Please number the boxes according to the usefulness to you of each type of application:

very useful | *2* | *moderately useful* | 1 | *not useful* | 0 |

Broadly speaking, queries can be about (A) words themselves, (B) their sense(s), and (C) related terms:

Queries about words themselves

Historical novel-writing

(a) What interactions were in common use in the period 1670- 1720?

□

English language teaching

(b) List current idiomatic phrases incorporating adverbs *not* ending in *-ly.*

□

(c) Most adjectives of the type _________can form adverbs/nouns ending in _________ . Are there any adjectives which do not appear to have such corresponding derivatives?

□

Literary criticism

(d) List the participial adjectives that end in *-ate,* (i.e. modelled on the Latin use). Which of them are obsolete or archaic? What is their date range? Who first used each of them?

|A: *Immolate,* obs. or arch., 1534-1830, More
Immurate, obs., 1593, R Barnes
Impersonate, 1820-67, Keats . . .|

□

Linguistics

(e) What is the distribution of *-lighted* as against *-lit* when the second element of a compound (e.g. *striplighted, -lit?* List all headwords (main and subordinate) with either form.

I I

(f) How many, and what kind of, words were borrowed in each of the periods specified from __________ (language specified)?

I I

History of science

(g) List all mineral names with the dates when they were named.

I I

Medicine

(h) List all medical syndromes, signs, etc. with the date at which they were named and the publication where the first description appears.

I I

Botany/Dialect studies

(i) List the vernacular English names for the plants of the order *Compositae.*

I I

History

(j) In an old manuscript document there are several illegible words. Each could be one of a number of possible combinations of letters and each combination could be an obsolete variant spelling of a standard word. Try all possibilities against *OED's* variant lists.

I I

Phonetics

(k) What words end with (contain) the sequence of sounds /i:tist/?
|A: defeatist, . . . elitist, . . . neatest, . . . retreatist, . . .|

I I

(I) What acronyms are treated like normal words when pronounced -- rather than as a series of letters (i.e. like *UNESCO* as opposed to *BBC)*

I I

Senses of words

Literary criticism

(m) What meanings of words does Milton follow Spenser in using?

I I

(n) What metaphorical senses of words, as evidenced by labels such as *transf., fig., concr.,* etc., does _______ |a writer| use, from evidence in the OED?

| |

Law/History

(o) A researcher into penal history needs to know what senses of words are given definitions containing the words: 'Instrument of punishment/of torture/of execution'.

|A: Halifax gibbet, . . . jougs, . . . maiden, . . . Scavenger's daughter. . . |

| |

Law/Pharmacology

(p) What is the field of meaning covered by the word *drug* (the main entry does not cover enough ground):
e.g. (1) What senses of words are defined using the word *drug?*
(2) What quotations are there later than 1884 containing *drug, narcotic, intoxicant,* [etc.]?

| |

Law

(q) What meanings of words in the dictionary are illustrated by a quotation from an Act of Parliament?

| |

Chemistry

(r) What terms have definitions that contain the character string . . *polymer* . . and what chemical structure does each definition give?

| |

Related terms

History I Politics

(s) Are there quotations of the period ________ ------------- that mention the place/institution/person _________________________ ?

|E.g. Q: 1500-1600; Mozambique? A: 1588 s.v. Inhabitance 1.|

| |

(t) Build up a profile of English knowledge of _______ | a place| over the period 1500-1800 from terms relating to that place.

| |

Specialist/Technical subjects

(u) How do the coiners of terms in |a specified subject| define their terms? Display the earliest quotation for each term in this subject, where the etymology does not give a reference to a coinage elsewhere than in the entry.

| |

(v) In what kinds of periodical are terms from statistics |or another specified subject| used?

|A: chi-square: *British Birds . . .*,
population^{2} (2d): *Biometrika, Jrnl. Gen. Psychol., Accountants Weekly*