The Digital Dictionary

Peter A. Stokes

Anglo-Saxonists have always been well represented in the field of Digital Humanities,[1] and perhaps the foremost among these has been *The Dictionary of Old English*. As we use the *Dictionary* and *Corpus* today, however, with their impressive modern interfaces and rapid search facilities, it is easy to forget that this project was first conceived in the 1960s when computing was paid for by the hour and the cutting edge in data storage was reel-to-reel magnetic tape. Despite these and other significant limitations, the Dictionary team chose to use computing technology from the very start, producing both the corpus and the dictionary itself in digital form, and they have managed to sustain this over some forty years. This achievement is a significant one, particularly as concerns about longevity of digital resources are still current, and so the lessons learned in this project are relevant to many of us now. These lessons are the ultimate subject of this paper, which will begin by considering the Dictionary of Old English Project and its development in the context of computing and digital humanities before discussing some uses and limitations of the *Dictionary* and *Corpus* and finally noting some brief lessons for large digital projects in general.

1 See, for example, the number of Anglo-Saxon projects and Anglo-Saxonists listed on websites such as *Digital Medievalist*, *Intute*, the *Old English Newsletter*, and the Department of Digital Humanities at King's College London at <http://www.digitalmedievalist.org/>, <http://www.intute.ac.uk/>, <http://www.oenewsletter.org/OEN/>, and <http://www.cch.kcl.ac.uk/research/projects/>, respectively.

A Brief History of the "Digital Dictionary"

The earliest published plans for the *Dictionary* are the transcripts of a conference held in Toronto on 21-22 March 1969; these were published as *Computers and Old English Concordances*, and while concordances were one of the two main topics of discussion, the other was "an exploration of the possibilities for beginning work on a large-scale Old English dictionary."[2] As well as being an important record for historians of computing in the humanities, this volume notes several issues that would arise again and again throughout the next forty years of the Dictionary of Old English Project, the consideration of which, in 1969, would have far-reaching consequences. These included basing the *Dictionary* on a digital corpus, the use of editions or manuscripts for this corpus, problems in lemmatization, and the decision to produce the *Dictionary* in a digital environment.

By 1971, several meetings had taken place towards building a machine-readable corpus, although at this point it was still debated whether the text should be keypunched directly or first typed, using a special font, and then read into the machine using OCR.[3] Visits were made to other historical dictionaries which used computers, and even at this early date, mention was made of "editing and layout of the *Dictionary* by use of display screens."[4] In April 1971, a working meeting took place in Toronto "to discuss encoding instructions with members of the Computistics Committee," and the results were published two years later.[5]

The concordancing system described at this date was to be written in Fortran on a Univac 1108. The Univac 1108 computer was modular and extensible but could easily fill a room with its various components, including not only the processor (or processors) but also the memory, console, card reader and punch, and reel-to-reel magnetic tapes; the purchase and installation cost was well over a million US dollars, and the operating cost to enter a million words of text and generate a concordance was estimated at about $30,000 in 1973.[6] The decision to use Fortran as a programming

2 Cameron, Introduction to "What Computers Can Do in the Humanities," in Cameron, Frank, and Leyerle, eds., *Computers and Old English Concordances*, 3.

3 Leyerle, "'The Dictionary of Old English': A Progress Report," 282.

4 Leyerle, "'The Dictionary of Old English': A Progress Report," 282.

5 The meeting was described by Leyerle, "'The Dictionary of Old English': A Progress Report," 283, and published by Venezky, "Computational Aids to Dictionary Compilation."

6 Walker, *Typical UNIVAC 1108 Prices: 1968*; and Venezky, "Computational Aids to Dictionary Compilation," 319-20. For further description and photographs of the UNIVAC, see Sperry Rand, *Univac 1108 II*.

language, rather than machine assembly language, proved far-sighted. Software in machine assembly language was faster to write and faster to run than that in Fortran, and at a time when computer usage was billed by the minute, this was a very significant consideration. However, since assembly language is also specific to the particular computer for which it was written, any software written in that language cannot be moved from one type of machine to another but must be rewritten almost from scratch.[7] In contrast, one of the primary strengths of "high-level" programming languages like Fortran is that they operate more or less independently of the machine. In other words, software written in Fortran for one system could, in principle, be moved to another one with relatively little effort: specifically, Venezky estimated in 1973 that the cost of transferring between systems could be reduced to a third of the initial development cost.[8] The decision to use such a high-level language, despite the greater initial cost in both time and money, proved valuable in the long term.

Richard Venezky and his team found that existing systems for text analysis were prohibitively expensive, and so by 1975 they had developed their own, LEXICO, written in Fortran and run on a Univac 1110: this offered the editors facilities for storage, editing, concordancing, and lemmatizing.[9] By the following year, the corpus had been converted to machine-readable form by typing the text with a customized typeball and then optically scanning the result. The alternative method was to punch the entire corpus onto cards and enter these into the machine, but this process would have been very long and error-prone. Since the maximum amount of data that could be entered onto an eighty-column card was eighty characters, the estimated thirty million characters of the corpus would have filled about 375,000 cards. In practice, very many more would have been needed, and, indeed, early discussions suggested a number in the range of about one million cards requiring a five-story basement to store them all; by comparison, Busa compiled some 800,000 cards for his *Index Thomisticus*.[10] If the

7 As one example, Paul Pillsbury noted that the University of Michigan's upgrade from an IBM 7090 to an IBM 360 in 1967 "necessitated a complete program translation" from one system to the other, a process that apparently began in early fall and was not finished until Christmas; see Pillsbury, speaking on "A Concordance to *The West Saxon Gospels*," in Cameron, Frank, and Leyerle, eds., *Computers and Old English Concordances*, 49.

8 Venezky, "Computational Aids to Dictionary Compilation," 319-20.

9 Venezky, "Unseen Users, Unknown Systems," 286-88; and Venezky et al., *Man-Machine Integration in a Lexical Processing System*.

10 For the early discussion, see Cameron, Frank, and Leyerle, eds., *Computers and Old English Concordances*, 12-13; for the *Index*, see Busa, "The Annals of Humanities Computing: The Index Thomisticus," 85. Busa's project is discussed further on p. 46, below.

cards were to become damp or were damaged, they could misfeed and would need to be fed into the reader by hand; moreover, if a box of cards were accidentally dropped, the cards would have to be re-sorted manually.[11] In contrast, a 2400-foot roll of magnetic tape in 1969 could hold as much data as up to 200,000 punched cards and cost about $16 — although tape had problems of its own, particularly the risk of inadvertent erasure.[12] A customized typeball was necessary to print the special characters which were not part of the standard English alphabet and so were not normally available for use; they could be provided relatively cheaply, however, simply by physically grinding an unwanted character off the ball and gluing a new one in its place.[13] After being scanned, the text was proofread and corrected, then separate concordances were generated for each text, and finally all the concordances were combined; all of this was done with the LEXICO system.

In 1976, about the same time that LEXICO became available, Bratley and Lusignan published a very perceptive discussion of the needs of the Dictionary's editors and the strengths and weaknesses of computers in serving these needs. Recognizing the limitations of computers, including that microfiche or even pencil and paper were more suitable tools for some tasks, they raised points which are still sometimes forgotten today, including that editors must be familiar with all of their material before they can make effective use of computer concordances. They also proposed a computer architecture for use by the editors which involved a mini-computer with local storage for articles being edited, connected to a remote computer centre which contained the entire corpus; this was partly a response to the difficulties and cost of storage on magnetic tape, and partly a way to isolate the Dictionary system from the central one, so that the Dictionary's computers could stay more or less constant even if the set-up at the remote computer centre was changed.[15]

---

11 Pierre R. Ducretet, speaking on "Computers and Literary Studies: Another View," in Cameron, Frank, and Leyerle, eds., *Computers and Old English Concordances*, 13-17; Burton, "Automated Concordances and Word Indexes," 139-40.

12 Ducretet, "Computers and Literary Studies: Another View," 12-13 and 17, and Venezky, "Computer Processing of Old English Texts," 67, both in Cameron, Frank, and Leyerle, eds., *Computers and Old English Concordances*.

13 Jess B. Bessinger, speaking on "A Concordance to *Beowulf*," in Cameron, Frank, and Leyerle, eds., *Computers and Old English Concordances*, 42-43.

14 Fred Robinson, review of *A Microfiche Concordance to Old English*, 133; Venezky, "Unseen Users, Unknown Systems," 286-88.

15 Bratley and Lusignan, "Information Processing in Dictionary Making," 142.

Despite the insights shown by Bratley and Lusignan, their system was never put into practice. Nor were several other possibilities which were considered in 1982 but all rejected: one system could not display pages of the final dictionary; another was a "turnkey" system, providing a complete fixed package which could not be re-programmed or customized sufficiently; and in yet another, the displays were too small and the text-editing functionality too poor.[16] Nevertheless, by 1985, the editors were in a position to describe what they saw as their "final" computer system. This was a hybrid of some of the earlier proposals and incorporated a "glorified file-server," a print server, and a workstation.[17] This workstation was itself innovative, showcasing inventions and approaches developed by Xerox and now considered standard, including the desktop metaphor, graphical icons, the mouse, and the so-called "what you see is what you get" or "WYSIWYG" interface, in which the screen directly reflects the final printed product with all its formatting. This last feature was particularly important to the editors of the Dictionary because of the typographical complexity of their work in both formatting and special characters and had been mentioned as an ideal as early as 1969.[18] Indeed, the novelty of such a sophisticated system is reflected in Toni Healey's description of it, published in 1985, in which she carefully described such processes as highlighting a word with a mouse and then selecting a format (such as bold) from a menu: all things which do not warrant mention in 2011.[19]

The *Microfiche Concordance to Old English* was published from the LEXICO system in 1980, the *Dictionary of Old English Corpus in Electronic Form* in 1981 (distributed on magnetic tape),[20] and *A Microfiche Concordance to Old English: The High-Frequency Words* in 1985. With these publications out of the way, the focus then shifted to the *Dictionary* itself. The first fascicle of the *Dictionary* was published on microfiche in 1986, a year after Healey's discussion of the computing system, and the second in 1988. The initial system described by Healey had only one workstation,

16 Healey, "The Dictionary of Old English and the Final Design of Its Computer System," 246-47. These systems were the VAX 750, a Xerox STAR network, and the Apple Lisa, respectively.

17 Specifically, the file server was a VAX 11/730 to run VMS and be programmed in C and assembler language; the print server was to be a Xerox 8044 (but was in fact an 8045: Healey, e-mail message to author, 12 Sept. 2009), and the workstation a Xerox STAR 1108 "Dandelion." Healey, "The Dictionary of Old English and the Final Design of its Computer System," 247-48.

18 Robinson and Bailey, speaking on "Concordances and Dictionaries," in Cameron, Frank, and Leyerle, eds., *Computers and Old English Concordances*, 99.

19 Healey, "The Dictionary of Old English and the Final Design of its Computer System," 248.

20 Healey, "The Dictionary of Old English: From Manuscripts to Megabytes," 157.

allowing only one person at a time to work at the computer while the other editors had to work on paper and then enter their notes into the machine. In a publication of 1988, Venezky noted that this system had been in use for about a year and a half at that time and that Xerox had donated four further workstations and a file server which allowed local storage of not only the entire corpus but also the complete concordance, lists of short-title references, headwords and frequency lists, and the index to Old English word-studies.[21] Venezky also described a "lexicographer's desktop" which he was completing at the time of writing: this was a graphical interface designed to "duplicate visually and functionally the work space that lexicographers normally adopt when working with books, card slips, paper and pencil."[22] By this time, the graphical user interface was better known — Venezky refers specifically to the "editing conveniences popularized by the Macintosh" — but such a system was still by no means universal.[23] In the same 1988 paper, Venezky also described a "portable editor," namely, a small and relatively inexpensive system which could be sent to editors outside Toronto, along with all the slip images for a given headword on disk, to allow "sense categories to be entered and edited and slips to be assigned to senses."[24]

In 1990, Lou Burnard, founder and director of the Oxford Text Archive, converted the *Electronic Corpus* to SGML and made it compliant with the Text Encoding Initiative's standard for scholarly data, thereby rendering it usable in principle on any machine with appropriate software.[25] The *Corpus* was published on disks in 1993, rather than the magnetic tape which had been used previously, and in the autumn of 1994 it became the single most requested corpus in the Oxford Text Archive for the tenth consecutive year.[26] In the following year, it was updated and converted to the latest version of the TEI standard (P3); it then filled nine 3.5-inch or eleven 5.25-inch disks. The *Corpus* was published on the World Wide Web in 1997,

21 Venezky, "Unseen Users, Unknown Systems," 288.

22 Venezky, "Unseen Users, Unknown Systems," 288.

23 Venezky, "Unseen Users, Unknown Systems," 288. For comparison, the first version of Microsoft Windows to be widely used on PCs was released two years later in 1990. Microsoft, *Windows History*, 3.

24 Venezky, "Unseen Users, Unknown Systems," 289. This system incorporated a Zenith Z-286 workstation running SCO, an operating system "generally compatible with" Berkeley 4.2 Unix; the editor used here was *vi*, an editor which is entirely text-based but which is still widely used and is included in Unix-based systems (including Apple's Mac OS X) today.

25 Healey, "Wood-Gatherers and Cottage-Builders," 36.

26 Healey, "Wood-Gatherers and Cottage-Builders," 36.

after an initial trial restricted to the University of Toronto.[27] At this time, the editors had two primary tools which had been custom developed: the "Corpus Browser" ("a user-friendly menu-driven program which allows a user to find specific citations or texts in the Corpus") and the "Text Analysis Language Browser" (allowing the editors "to search both Latin and Old English with Boolean searches and searches on regular expressions").[28]

Although Venezky had foreseen publication of the *Corpus* on CD-ROM by 1988,[29] this was not to happen for another twelve years, when the SGML, HTML, and XML versions were all published on a single disk in the year 2000, and revised versions were issued in 2004 and 2009. By 2002, the online *Corpus* allowed a range of possible searches, including parts of words, beginnings of words, simple phrases, Latin text, and Boolean operators. Searches could also be restricted by the short-title or the "Cameron Number," a unique number assigned to each text.[30] This very simple feature allows a very wide range of different searches, since Cameron's system of numbering encodes much information about the text. Thus, the Cameron Numbers for poetic texts begin with "A," for prose texts with "B," and so on; hence, a search of all poetic texts can be achieved by limiting to Cameron Numbers beginning with "A." Similarly, poems from the Junius manuscript, for example, have the Cameron Number "A1," and thus all the poetry from that particular manuscript can be searched in the same way.[31] The problem with searching the *Corpus* was still the wide range of spellings possible in Old English, and unfortunately the text was still not lemmatized, nor was it possible to use wildcard or regular expression searches; the latter restriction was due to the limitations of *xpat*, the OpenText software which was used to develop the Web *Corpus*.[32] To overcome this problem, a "Variant Spellings" tool and

27 Healey, "Wood-Gatherers and Cottage-Builders," 37; Healey, "The Dictionary of Old English: From Manuscripts to Megabytes," 158.

28 Healey, "Wood-Gatherers and Cottage-Builders," 38.

29 Venezky, "Unseen Users, Unknown Systems," 290.

30 For a full list of Cameron Numbers, see Cameron, "A List of Old English Texts"; Cameron et al., *Old English Word Studies*.

31 However, the *Corpus* cannot generally be used to study scribal habits or represent the manuscript (*pace* Healey, "The Dictionary of Old English: From Manuscripts to Megabytes," 161; and Drout et al., "Lexomics for Anglo-Saxon Literature" [2009 conference paper and p. 4 of the 2010 article of the same title]), since the texts are edited not diplomatic, and texts from the same manuscript often come from different editions following different editorial practice. For further discussion, see below, pp. 54-56.

32 Healey, e-mail message to author, 12 Sept. 2009.

a "Word Wheel" have been provided for the Web version of the *Corpus*, the first of which allows searches with *eth* and *thorn* combined and with wildcards for vowels, and the second allows users to select from a list of all words in the *Corpus* rather than having to guess spellings; the latter also provides a direct link to the search engine on the Web *Corpus*, permitting rapid access to the citations. The latest (November 2009) release of the *Corpus* includes further corrections to the texts, permits "Simple," "Boolean," "Proximity," and "Bibliography" searches, and provides Old English, Latin, Greek, and Runic "Word Wheels."

The *Dictionary* itself was first published on CD-ROM in 2003 with the letters *A* to *F*. Like the *Corpus*, the *Dictionary* included full content in HTML, SGML, and XML, as well as the XSLT required to generate the HTML from the XML. This publication also featured electronic texts in structural markup rather than the previous typographical one.[33] In essence, the initial typographical markup reflected what the dictionary text should look like when printed — which words should be printed in bold, which in italic, and so forth. Although useful for a print publication, this sort of markup is not very helpful in an electronic environment, where it is much more useful to encode the function of the words, labelling which are headwords, which definitions, and so on. This means, first, that separate rules can be defined to govern how each function should be displayed (all headwords in bold, for example), resulting in formatting which is more consistent and less prone to human error, and also much easier to change if necessary. The second and ultimately greater advantage, however, is that this knowledge of the text's structure can be used by the computer. One can now run searches against particular fields, looking only for headwords, or only for attested spellings, and so forth. The text can also be published in a variety of formats with minimal effort by the editors, and Healey has already suggested several such possibilities.[34]

The CD-ROM version of the *Dictionary* includes "DOESearch," an interface which allows searching of the different fields as just described, as well as "regular expression" searches which include fairly complex wildcards. Unfortunately, the interface operates on Microsoft Windows only, but the basic content can be viewed in any Web browser, and the browser form includes navigation by headword. Since the

33 Healey, "The Dictionary of Old English: From Manuscripts to Megabytes," 171-77; Healey, "The *Dictionary of Old English*: The Next Generation(s)," 293.

34 Healey, "The Dictionary of Old English and the Final Design of its Computer System," 246.

SGML and XML are also included on the CD-ROM, users with the requisite skills can very easily create interfaces of their own. The markup is based on "[t]he field structure of a DOE entry" but is apparently a custom-designed schema and is not compliant with the TEI standard for dictionaries;[35] the Project staff does plan to re-examine TEI under the new P5 standard for dictionaries as they aim to conform to this wherever possible.[36] The CD-ROM *Dictionary* was further extended in 2008, in content with the inclusion of *G* and in interface with the addition of Boolean searches across fields, links to the full *Oxford English Dictionary Online*, and a short-title list and bibliography of Latin texts.[37] In much the same way, a web-based version of the *Dictionary* was published for *A* to *F* in 2003 and for *A* to *G* in 2007. The *A* to *G* version also includes simple and Boolean searches across different fields, Old English and Latin short-titles and bibliography, and links to the *Oxford English Dictionary Online*; it therefore provides the same functionality as the CD-ROM version but does not depend on a Windows environment, although it does not include access to the SGML or XML files.

The fascicle for *F* was also published on microfiche in 2004, and that for *G* in 2008. Healey has explained the continuing use of microfiche as arising from the wish to provide these new resources to colleagues in "incredibly challenging circumstances" who need the material in as cheap and accessible a format as possible, but she has also stated in the same context that the Project staff expect to stop providing this "with the next letter or two" as access to technology spreads.[38] On the other hand, microfiche does not suffer from the same problems concerning long-term sustainability which affect digital resources, and for this reason such "hard copies" are often still favoured by libraries and, indeed, by scholars; staff at the Dictionary of Old English Project therefore plan to continue publishing on microfiche until these issues are resolved.[39]

35 Healey, "The *Dictionary of Old English*: The Next Generation(s)," 293. TEI Consortium, *TEI P5: Guidelines*, Ch. 9.

36 Healey, e-mail message to author, 12 Sept. 2009.

37 However, a separate subscription is required to view content in the *Oxford English Dictionary Online*, and this subscription is not included with any version of the *Dictionary of Old English*.

38 Healey, "The *Dictionary of Old English*: The Next Generation(s)," 304.

39 Healey, e-mail message to author, 12 Sept. 2009.

Related Projects

It is worth comparing the history of the *Dictionary* to that of some related projects. At the time of the first meeting in 1969, electronic editing and the use of an electronic corpus and concordance were unusual, although not unprecedented.[40] Indeed, the pioneering work of this sort is probably the *Index Thomisticus*, a full electronic concordance of all the works of Thomas Aquinas, which was started in 1946 by Father Roberto Busa and completed with a print publication in 1974-1980 (comprising fifty-six volumes), a CD-ROM in 1992, and a web-based version in 2005. Busa's work was exceptional, however: he had already published a machine-generated concordance in 1951, despite a report from IBM stating that such work was beyond the capability of their machines, and the Association for Literary and Linguistic Computing and Association for Computing in the Humanities (ALLC/ACH) founded a prize in his honour in 1998, of which he was the first recipient.[41] Apart from the *Index*, computers were being used for several lexicographical projects when the *Dictionary of Old English* was first planned: these include the *Dictionary of American Regional English* (*DARE*) and four historical dictionaries noted by Leyerle, namely, "the French *Trésor* at Nancy, the *Italian Historical Dictionary* at Florence, the *Historical Dictionary of Hebrew* at Jerusalem, and the *Dictionary of the Older Scottish Tongue* [*DOST*] at Edinburgh."[42] In an article published as these meetings were taking place, Richard Bailey listed his *Early Modern English Dictionary* as a further example and provided a fairly extensive bibliography of these and other lexicographical projects which used computers.[43] However, the ways in which the computers were used varied widely. Staff at *DARE*, for example, foreground their use of computers for statistical analysis and distribution maps rather than citations.[44] The *Early Modern English Dictionary* was apparently planned to be built from an electronic corpus like that of the *DOE*, and Bailey et al. published the corpus on microfilm in 1975 and electronically in 1996, but the dictionary seems to

40 Burton, "Automated Concordances and Word Indexes"; Hockey, "The History of Humanities Computing," 4-7.

41 Busa, "The Annals of Humanities Computing," 84. Hockey, "The History of Humanities Computing," 4.

42 Leyerle, "'The Dictionary of Old English': A Progress Report," 282-83.

43 Bailey, "Research Dictionaries," 170, n. 3.

44 *DARE* [n.d.] "About DARE" – "History."

have been abandoned. Similarly, the *Historical Dictionary of Hebrew* is based on a digital corpus and concordance which was begun in 1964 and published on CD-ROM first in 1989 and online first in 2006. Unlike the corpus used by the Dictionary of Old English Project, the Hebrew corpus had to be prepared directly from manuscripts rather than editions, and is also fully lemmatized, but this corpus is still not complete, and no fascicles of the dictionary itself have been published.[45] A similar pattern of publication-media is evident with the *Trésor de la Langue Française*, published first on CD-ROM and then online. The first volume of *DOST* was published in 1931, well before the digital electronic computer was invented, and subsequent volumes were published in print only, but the content was later incorporated into the online *Dictionary of the Scots Language*.[46]

Although a full comparison of these projects is beyond the scope of this paper, it is evident that the *Dictionary of Old English* is unusual in several respects. It was one of only a very small number to use electronic corpora at all when it was begun. It is even more unusual in its basis in such a corpus (rather than being a reworking of existing dictionaries), as well as in its compliance with standards (at least for the *Corpus*) for making the entire corpus easily available in "human-friendly" format with search interface and so on while also supplying the "machine-friendly" data underneath (now XML). This concern has been evident from the start, and even in 1969 a significant part of the discussion was about standards for machine-readable texts, in terms of character encoding, text encoding, and storage formats.[47] The importance and value of the *Corpus* in particular has been very much increased by this aspect of planning the *Dictionary* and its related resources.

45 Merkin, Busharia, and Meir, "The Historical Dictionary of the Hebrew Language"; Mishor, "The Philological Treatment of Ancient Texts"; *Ma'agarim: A Database*; and *Ma'agarim: Online Historical Dictionary of Hebrew*.

46 Rennie et al., eds., "About the DSL," *Dictionary of the Scots Language*.

47 Venezky, "Concordances to the Ruthworth [*sic*] Matthew and the Vercelli Homilies," in Cameron, Frank, and Leyerle, eds., *Computers and Old English Concordances*, 65-81 and 111; see also Venezky, "Computational Aids to Dictionary Compilation," 321-23; Bratley and Lusignan, "Information Processing in Dictionary Making," 139; Healey, "Wood-Gatherers and Cottage-Builders," 36-37; Healey, "The Dictionary of Old English: From Manuscripts to Megabytes," 170; and Healey, "The *Dictionary of Old English*: The Next Generation(s)," 290-92.

Reception and Impact: Uses and Limitations of the *Dictionary* and *Corpus*

Use and Re-use: The DOE *and Digital Standards*

One direct benefit of the Dictionary of Old English Project's compliance to standards is longevity. Very much effort has been devoted to converting the *Corpus* and *Dictionary* to different formats throughout the project's life as new standards have emerged; nevertheless, these conversions have at least been possible and completed, and thus the material has been continuously usable for almost thirty years since its first publication in 1981. This compliance with standards in the project's content has also aided maintenance of the software used to process it. As discussed above, the decision to write the first versions of this software in high-level languages such as Fortran made the software much more portable across different systems; similarly, as standards emerged for encoding content, so the content became less and less dependent on any one custom-built and machine-dependent piece of software.

As a result of this flexibility and longevity, the Dictionary of Old English Project stands in sharp contrast to so many other digital projects which have disappeared or become unusable within a decade or less.[48] Indeed, just as the sustainability of digital projects remains an issue today, and as it discourages some from undertaking digital publication at all, the Dictionary project is a good example not only of the difficulties in maintaining such a resource but also of the ways in which these difficulties can be overcome with sufficient care and planning. It is unfortunate in this context that the *Dictionary* itself, while marked up in XML, is not yet compliant with the TEI standard. This contrasts with past practice (as demonstrated by the fully-compliant *Corpus*) and is presumably in part a result of the cost and effort required for conversion but also of the TEI Guidelines being insufficient for the Dictionary's needs at the time.[49]

---

48 Perhaps the best-known of these is the *Domesday Project*, which was completed at the end of the 1980s and distributed on twelve-inch optical disks in a proprietary format which was unusable fifteen years or so later; see O'Donnell, "The Doomsday Machine," and O'Donnell, "Disciplinary Impact and Technological Obsolescence," 67-68. Even resources developed in SGML can rarely be used to their full potential today, since SGML browsers are normally proprietary, and conversion from SGML to XML is often difficult; for one of many examples, see O'Donnell, *Cædmon's Hymn: A Multimedia Study*, with brief observations by Stokes, Review of *Cædmon's Hymn*, and further discussion by O'Donnell, "Disciplinary Impact and Technological Obsolescence," 71.

49 Healey, "Wood-Gatherers and Cottage-Builders," 36 and 41-42. Healey, e-mail message to author, 12 Sept. 2009.

Indeed, the authors of the TEI Guidelines themselves acknowledge the difficulties in developing a general system for markup of dictionaries, citing a fairly extensive discussion in the literature, and they note the probable need to expand the schema in future, but in practice this chapter of the Guidelines has had little expansion in recent versions.[50] As a result, customization of the standard may well be necessary: such customization is entirely within — and expected by — the standard,[51] but it can require careful analysis and a high level of expertise, along with the time and effort that this entails. Nevertheless, such additional investment in standards compliance has paid off in the past and is sure to do so again in future.

As well as helping to ensure longevity, such compliance to standards has also made it relatively easy to re-use the data in other applications. Indeed, the Dictionary staff have always been eager to share their data as much as possible, and they have always planned for one outcome of their work to be a distributed database for research.[52] This enlightened policy has had several benefits. First, it means that the entire evidence-base for the Project is available for scholarly scrutiny, an admirable policy which has been advocated in the Humanities but rarely followed in practice.[53] Furthermore, electronic resources are routinely used in ways that the compilers did not anticipate, and the more standards-compliant the resource is, and the more the underlying data is revealed, the more freedom scholars have for re-using that material. The staff at Toronto have frequently expressed surprise at the ways in which their material has been used and at the changes in this usage. For example, during the 1980s, demand for the pre-built concordances was much greater than for the *Corpus* itself, since the computing required to use the corpus was expensive, (relatively) laborious, and beyond the capacity of the "lone scholar" sitting at his or her desk.[54] However, this soon changed, and by the mid to late 1990s anyone who had the most basic

50 TEI Consortium, *TEI P5: Guidelines*, §9. For development of the *Guidelines*, compare TEI Consortium, *TEI P4: Guidelines*, §12; and TEI Consortium, *Guidelines . . . (TEI P3)*, §12.

51 TEI Consortium, *TEI P5: Guidelines*, §23.2.

52 Venezky, "Computational Aids to Dictionary Compilation," 311; Healey, "Wood-Gatherers and Cottage-Builders," 37-38.

53 Bailey, "Research Dictionaries," 171-72; Jenkyns, "The Toronto *Dictionary of Old English* Resources: A User's View," 390; de Schryver, "Lexicographers' Dreams," 167-71; and Stokes, Review of *Cædmon's Hymn*, §10.

54 Venezky, "Unseen Users, Unknown Systems," 288; and Healey, "Wood-Gatherers and Cottage-Builders," 35.

desktop computer could not only access the *Corpus* but also run software to generate concordances, "reverse" concordances, frequency-lists, and much more.[55] As a result, the *Corpus* has had a range of other uses, including incorporation into the much larger Helsinki corpus and related work and into a syntactically annotated corpus of Old English prose; it has also been used to study the politics of Old English language change or teach style in *Beowulf*, to identify manuscript fragments, and to study unassimilated Latin words in Old English, and it has served as the basis for an extensive online resource of Anglo-Saxon charter bounds and has been employed in a pilot study of authorship attribution, amongst others.[56] For several reasons, this process of re-use is very much easier than it might have been, in particular because the *Corpus* is TEI-compliant and because all of the underlying SGML and XML is made available, something that is not done in many new projects even today. In 1992, for example, the *Corpus* was "singled out as 'unique'" by the University of Michigan as the only one which could be included in their system without modification because of its openness and compliance to standards.[57]

The integration and re-use discussed so far concerns the *Corpus* rather than the *Dictionary*. However, the electronic *Dictionary* can also be integrated into other projects, albeit in a different way, in a manner envisaged some time ago by several commentators. Joy Jenkyns, following Joachim Neuhaus and writing just before the

55 Healey, "Wood-Gatherers and Cottage-Builders," 35. Some examples of freely available software to perform such analyses are Concordance, XAIRA, PhiloLogic, and TextSTAT. Concordance is a basic tool for Windows which is relatively easy to use but works only on plain text. XAIRA and PhiloLogic are much more complex and powerful, and can take advantage of texts marked up with TEI-compliant XML for lemmas, part-of-speech tagging, and so forth. TextSTAT falls in between: it can accommodate plain text or HTML but can also be modified relatively easily for XML-encoded texts; see Stokes and Pierazzo, "Encoding the Language of Landscape," 226-27. The tools are available at <http://www.concordancesoftware.co.uk/>, <http://www.oucs.ox.ac.uk/rts/xaira/>, <http://philologic.uchicago.edu/>, and <http://neon.niederlandistik.fu-berlin.de/textstat/>, respectively.

56 These uses are discussed, respectively, in Healey, "Wood-Gatherers and Cottage-Builders," 37-38, and Healey, "The Dictionary of Old English: From Manuscripts to Megabytes," 170-71; Taylor et al., eds., *The York-Toronto-Helsinki Parsed Corpus*; Price-Wilkin, "A Campus-Wide Textual Analysis Server," §§4.1-4.2; Healey, "The Dictionary of Old English: From Manuscripts to Megabytes," 170; *LangScape* <http://www.langscape.org.uk>; and Drout et al., "Lexomics for Anglo-Saxon Literature," and below, pp. 54-56.

57 Healey, "Wood-Gatherers and Cottage-Builders," 36, citing Price-Wilkin, "A Campus-Wide Textual Analysis Server," 127 §2.2.

public advent of the World Wide Web,[58] suggested electronic links between text, manuscript image, variant manuscripts, dictionary entries, concordances, and other lexicographical resources such as the electronic *Thesaurus of Old English*.[59] Similar ideas have also been expressed by Healey[60] and have been partially implemented in the latest release of the *Dictionary* by including hyperlinks to corresponding headwords in the *Oxford English Dictionary Online*. To date, these links are relatively "shallow," in that they simply allow readers to move from an entry in the *DOE* to the corresponding entry in the *OED*, rather than "deep" integration where the underlying content (data and markup) is used rather than the interface alone. Even these "shallow" links are useful, however, and a fully linked version of the sort that Jenkyns and Healey have both described would be extremely valuable, with two-way connections not only between the *Thesaurus*, *Corpus*, and *Dictionary*, but also to other resources like the online *Dictionary of Medieval Latin from Celtic Sources*, the *Perseus Project*, digital editions, and digitized manuscripts. Unfortunately, the effort required to do this is well beyond the immediate scope of the *Dictionary*, and simply creating the links is by no means trivial. Not only does it require a good understanding of the file-structure and organization of the other projects, but it raises significant problems of sustainability. In addition to the usual problems of sustaining the *Dictionary* itself, the links also depend on the other projects both surviving and retaining their overall structure. If one resource were to change its system of internal URLs, for example, then all the links to that resource would be broken and would need to be changed. Linking from the *Dictionary* to the *Corpus* would presumably be straightforward, since both are maintained in one place and because there is a direct, one-to-one link from *Dictionary* citation to *Corpus* text which must have already been identified during the *Dictionary*'s compilation. Linking from the *Corpus* to the *Dictionary*, however, or linking to or from manuscript images, is a much more complex procedure. Despite these challenges related to the *Dictionary*, it is worth noting that the projects mentioned above have already demonstrated the value of "deep" integration of the *Corpus*, and this suggests that such work with the *Dictionary* will also bring returns.

58 For the date of which, see Cailliau, *WWW Project History*.

59 Jenkyns, "The Toronto *Dictionary of Old English* Resources: A User's View," 414, citing Neuhaus, "Design Options for a Lexical Database of Old English."

60 Healey, "The Dictionary of Old English: From Manuscripts to Megabytes," 173-77.

Limitations: Diplomatic vs. Edited Corpus

One of the difficulties with the re-use of such data is that one must understand precisely the nature of the data being re-used. For example, all texts are edited according to different principles and assumptions, and these principles and assumptions will dictate the way in which the texts can and cannot be used. A Lachmannian edition, for example, is very different from a diplomatic transcript — both are useful but for different ends — and even two diplomatic editions will inevitably follow different principles and produce slightly different texts, as no use of a manuscript can ever be entirely objective.[61] It is partly for this reason that any scholarly printed edition must contain an introduction in which the editorial principles are clearly stated. In digital texts, however, this statement of principles is often omitted or unclear, and this lack of clarity limits the degree to which these texts can be re-used. During the early stages of development of the *Corpus*, the important question was raised several times whether this resource would contain diplomatic texts from manuscripts, draw on existing editions, or combine these two options.[62] In the end, the practice seems to have been to draw on editions, but the Dictionary staff have microfilm copies of all manuscripts in Neil Ker's *Catalogue of Manuscripts Containing Anglo-Saxon*, and the *Corpus* does include corrections based on the manuscripts.[63] Furthermore, some texts have been emended to restore manuscript readings from the critical apparatus of the editions,

61 For further discussion on this point, see below, pp. 59-60 incl. n. 95.

62 Cameron, Frank, and Leyerle, eds., *Computers and Old English Concordances*, *passim*, e.g., Cross and Venezky, "Discussion," 46, and Walter Bak, speaking on "A Concordance to MS Hatton 20," 62-64; Gneuss, "Guide to the Editing and Preparation of Texts"; and Jenkyns, "The Toronto *Dictionary of Old English* Resources: A User's View," 381-82. Cameron, "A List of Old English Texts," also specifies one manuscript to be followed for each text.

63 A search of the current *Corpus* for texts containing the word "corrected" in their bibliographical statement returned sixty-six examples; this figure suggests significant but by no means universal correcting. As Jenkyns has noted, however, it is not always clear if or when texts have been corrected; Jenkyns, "The Toronto *Dictionary of Old English* Resources: A User's View," 385. Although the situation has somewhat improved since 1991, there are still discrepancies, and in particular the *Dictionary*'s "List of Texts Cited" does not specify which editions have been altered. For example, metadata in the *Corpus* text of Ælfric's *Hexameron* (*Corpus* short-title "ÆHex," Cameron Number B1.5.13) describes the source as "corrected against MS," but the same text in the *Dictionary*'s "List of Texts Cited" does not specify these corrections; similarly the form *asæled* in *Genesis A*, line 2196, is flagged as an altered form in the *Corpus* ("GenA,B," Cameron Number B1.1) but is listed without comment in the *Dictionary* (s.v. *a-sælan*). On the other hand, the metadata for *Genesis A* in the *Corpus* does not refer to any intervention by the staff in Toronto. Notes provided with the online *Dictionary of Old English: A to G* also state that "spellings not in the Corpus but found in variant

or at least to flag editorial interventions in the electronic text.[64] This policy gives a useful compromise between the enormous time required to transcribe texts anew and the problems of drawing on potentially unreliable editions. However, it is important for users to understand this distinction between manuscript and edition and to know which sources were used for which texts. As Jenkyns has commented, some of the texts in the *Corpus* come from very old editions, and some of these editors altered their texts almost arbitrarily from the manuscripts, particularly in orthography; not all of the manuscript forms have been restored or even noted by the staff in Toronto.[65] Thus, Jenkyns's sample of the charter bounds in the *Corpus* led her to conclude that they constitute "carefully copied but unrevised versions taken mainly from 19th-century publications":[66] those which come from Kemble's *Codex diplomaticus aevi Saxonici* (1839-1848), for example, show substitution throughout of *thorn* for *eth*, amongst other editorial interventions.[67] The charter bounds have generally been corrected since then on Jenkyns's advice,[68] and, indeed, Healey estimates that a great number of texts in the *Corpus* have had some sort of intervention by the project's staff.[69] Although many examples can easily be found of nineteenth-century editions without any note of correction in the accompanying metadata,[70] Healey also notes that the *Corpus* has been under a Revision Control System since 1998: all corrections are now recorded automatically, as is the name of the person making them.[71] This is an important record but one which, unfortunately, is not made readily available.

---

manuscripts are also included" under Attested Spellings, but these variants are not flagged in the dictionary entry, so neither their presence nor their source can be determined. Hereafter all texts in the *Corpus* will be referred to by their short-title and Cameron Number; for Cameron Numbers, see above, note 30.

64 Healey et al., "Documentation on SGML, XML, and HTML Corpus" 2 and 3, and in Corpus 2009 Release (CD-ROM version only).

65 Jenkyns, "The Toronto *Dictionary of Old English* Resources: A User's View," 384-85.

66 Jenkyns, "The Toronto *Dictionary of Old English* Resources: A User's View," 384.

67 Jenkyns, "The Toronto *Dictionary of Old English* Resources: A User's View," 392, n. 42.

68 See, for just one example, "Ch283" (B15.8.57) in the *Corpus*.

69 Healey, e-mail message to author, 12 Sept. 2009.

70 Just two sets of examples from the November 2004 release are Ælfric's *Lives of Saints* ("ÆLS," B1.3) and a number of homilies from Napier's edition of Wulfstan's homilies. Many of these texts came from later reprints, but there is no clear statement in the files for these texts whether the reprints themselves were ever corrected. The header in the 2004 release of "HomU 35.1" (B3.4.35.1) hints that at least some of these have been silently corrected ("slight variation [from the 1883 edition] occurs in lineation due to insertion of readings from MS. E"), but this note has been removed from the 2009 release.

71 Healey, e-mail message to author, 12 Sept. 2009.

The base editions are identified clearly in both the *Corpus* and *Dictionary*, and thus it is no fault of the Toronto editors if scholars misuse these resources, but one must be cautious of such misuses. Even Healey has written that the *Corpus* can be used to "study the scribal habits of a particular manuscript" by filtering on Cameron Number, and others have written of building "virtual manuscripts" or determining "the most common — or the rarest — words in [a particular] manuscript."[72] However, it must be remembered that this "virtual manuscript" may be a collection of texts drawn from different editions with different editorial policies and may therefore not reflect any given "real" manuscript. For this reason, the term "manuscript" is misleading at best in this context.

Particularly interesting in this regard is a new project designed to apply corpus linguistics and automatic authorship attribution to the *Corpus*.[73] Despite being in the early stages of development and still (by the researchers' own admission) using very crude methods, this approach has been applied successfully to analyses of single texts from single editions, such as automatically separating *Genesis A* from *Genesis B*, and the Old English *Azarius* from *Daniel*.[74] Effective as this has been, however, such an approach is very much harder to justify when applied to different texts from different editions and different manuscripts. For example, the same group has also considered Bede's account of the Life of St. Cuthbert in the Old English version of his *Historia Ecclesiastica* and has tried to automatically identify uses of this text in Ælfric's *Catholic Homily* II.10.[75] The latter text in the *Corpus* is Malcolm Godden's 1979 edition, and the former is the 1959 reprint of Thomas Miller's edition first published in 1890-1898.[76] The two editors were working nearly a century apart and unsurprisingly follow different editorial practices; they also base their texts on different manuscripts copied by different scribes who were writing different forms of Old English. Specifically, Godden's text is based on a single West Saxon manuscript, and emendations are limited to "[t]he few obvious errors not corrected by the scribe himself."[77]

72 Healey, "The Dictionary of Old English: From Manuscripts to Megabytes," 161-62. Drout, "Lexomics at Kalamazoo," *Wormtalk and Slugspeak*. See also LeBlanc et al., *Wheaton College Lexomics Group* – "Tools."

73 LeBlanc et al., *Wheaton College Lexomics Group*.

74 Drout et al., "Lexomics for Anglo-Saxon Literature" (2009 conference paper and 2010 publication).

75 Drout et al., "Lexomics for Anglo-Saxon Literature" (2009 conference paper and 2010 publication). This study of works by Bede and Ælfric was omitted from the 2010 publication.

76 "ÆCHom II, 10" (B1.2.11) and "Bede 4" (B9.6.6), respectively.

77 Godden, ed., *Ælfric's Catholic Homilies: The Second Series – Text*, xciv.

In contrast, Miller described his edition as "a 'contamination' of texts" based on four different Anglian manuscripts, having "discarded" the principal West Saxon witness,[78] although in practice he based his text on one principal manuscript. Fortunately, both editors also endeavoured to reproduce their base manuscripts accurately and to note editorial interventions, unlike some others in the *Corpus*.[79] Furthermore, although the *Corpus* metadata does not refer to emendation of either text by the staff at Toronto, comparison with the printed editions shows that Miller's alterations of the base witness have been flagged in the electronic text, and thus it is possible here at least to identify these cases, though not to restore the manuscript readings.[80] Even if these emended words were discarded, however, one would still expect the two editions to show different orthography, particularly since one presents an Anglian text copied in the first half of the tenth century and showing varying orthography between scribes and the other a West Saxon one from the very end of the tenth century and mainly copied by one scribe probably working very closely with the author.[81] However, the *Corpus* is not lemmatized and so the authorship attribution software cannot recognize the same word spelled in two different ways; in other words, the two texts may have a very large overlap in vocabulary but this overlap may be missed entirely by the software. Thus, it should not be surprising that the project team found that their software identified Ælfric as Ælfric and Bede as Bede rather than find one author's use of the other.[82] Nevertheless, I would suggest that even this statement is a misinterpretation: the software is not distinguishing Ælfric's writing from Bede's (nor even that of Bede's translator), but rather Godden's edition from Miller's, or at best the West Saxon copy

78 Miller, ed., *The Old English Version of Bede's Ecclesiastical History*, 1:v.

79 For examples see Jenkyns, "The Toronto *Dictionary of Old English* Resources: A User's View," 384-85.

80 At the time of writing, it appears that the *Old English Lexomics* data does not use this information but includes these editorial forms without comment. For example, a small sample of words flagged in the *Corpus* text of "Bede 4" (B9.6.6) gives *bodode*, *lyfesne*, and *towurpun*: these are all editorial reconstructions and are not found in the manuscript. However, the so-called "Virtual Manuscript" Tool lists all of them without comment (LeBlanc et al., *Wheaton College Lexomics Group* – "Tools" – "Virtual Manuscript" – "Text B09.006.006_Bede_4_T06900"). In this respect, we would all do well to heed Busa's warnings from nearly thirty years ago about the use of electronic texts without annotation; Busa, "The Annals of Humanities Computing," 86.

81 For the former, see Ker, *Catalogue of Manuscripts Containing Anglo-Saxon*, 428-29; Miller, ed., *The Old English Version of Bede's Ecclesiastical History*, 1: xiii-xv. For the latter, see Ker, *Catalogue*, 13 and 21; Godden, ed., *Ælfric's Catholic Homilies: The Second Series – Text*, xliii.

82 Drout, "Lexomics for Anglo-Saxon Literature" (2009 conference paper).

of Ælfric from the Anglian translation of Bede. It is probably identifying editors or scribes at least as much as authors.

This project has already made important findings and demonstrates the value of such studies for the *Corpus*, but its results must be interpreted carefully. In order to permit genuine studies of authorship across editions and manuscripts, the *Corpus* must first be fully lemmatized: once this is done, it will be an exceptionally powerful resource. In contrast, to use the *Corpus* for studies of scribal practice, we must first carefully examine the metadata for each text, comparing it with the manuscripts if necessary, to understand its individual status, specifically considering the reliability and editorial practices of the source and any subsequent correction or flagging in Toronto. This is already possible to some extent, if rather painstaking, with the existing data and metadata, but it would be very much easier if more extensive and precise information were provided, including (but not limited to) that described by Speirs.[83] Even this will be only a sample of extant manuscripts, however, with one for each text, and, thus, cannot be used to study Anglo-Saxon scribes in general; for that, one must turn to projects like the *ManCASS C11 Database of Scripts and Spellings* or the *LangScape* resource of Anglo-Saxon charter bounds, both of which were designed from the start to represent the orthography of all surviving manuscripts, and both of which contain very much less text than the *Corpus* and yet still took years of painstaking research to compile.[84]

Limitations: Lemmatization

Some of the difficulties outlined above would be ameliorated somewhat if the corpus were lemmatized. Indeed, the topic of lemmatization has come up several times already in this paper and has also arisen regularly in discussions of the *Dictionary* in general and of the *Corpus* in particular.[85] As these discussions show, it has always been a desideratum to lemmatize the *Corpus*, and tools to help do it have been

83 Speirs, "Lexicography and Corpus-Tagging," 137-42.

84 Scragg et al., *ManCASS C11 Database Project*. *LangScape: The Language of Landscape:* <http://www.langscape.org.uk>.

85 Cameron, Frank, and Leyerle, eds., *Computers and Old English Concordances*; Venezky, "Computational Aids to Dictionary Compilation," 316-19; Bratley and Lusignan, "Information Processing in Dictionary Making," 135; Venezky, "Unseen Users, Unknown Systems," 287; Fred Robinson, review of *A Microfiche Concordance to Old English*, 134; and Speirs, "Lexicography and Corpus-Tagging," 142-46.

developed or suggested regularly as well.[86] Apparently, the *Corpus* is being lemmatized as the staff at Toronto work through the alphabet,[87] presumably marking up only those words which are included in the fascicle being developed at that time, but the data is not available in the published *Corpus*. Unfortunately, to lemmatize any text is tedious and error-prone and can be done only partially automatically at best, and a historical corpus is generally much more difficult to lemmatize than a modern one. The first reason for this is varying orthography: in general, modern languages (and even Latin) have more or less fixed spelling, but this is certainly not true for Old English. Not only do scribes show little, if any, standardization in their practices, but there is also variation across geography (dialect) and time.[88] Indeed, even modern scholars do not always agree which dialect of Old English to use for dictionary headwords, let alone which spelling, with grammars and general-purpose dictionaries usually using West Saxon and place-name dictionaries using Anglian forms.[89] Furthermore, since Old English is much more inflected than Modern English, different endings and changes due to grammatical form must also be considered. As a result, lemmatizing a corpus of Old English requires matching a potentially very large range of attested spellings to a single headword, a process that can be automated in principle but only if all the possible spellings for each headword are known and have been fed into the computer in advance. This problem is further compounded by the problem of word-division: it can be difficult to decide what to treat as separate words and what as compounds, and an edition in the *Corpus* may well follow principles which are different from those of the *Dictionary*. In such a case, any automatic system for lemmatization would have to recognize the possible ways in which a given headword could be divided, and search the *Corpus* accordingly. The form *æsc-rind*, for example, is listed in the *Dictionary* as a compound. However, the *Corpus* could well contain a text with *æsc rind* given as two separate words: thus any automatic lemmatization must search for both *æscrind* and *æsc rind*. This may seem trivial, but if alternative spellings are considered, including different inflections of the words (the *Dictionary*, for example, lists *æscrind*, *æscrinde*, and *æscrinda* as attested spellings), then

86 Venezky, "Computational Aids to Dictionary Compilation," 317-19; and Venezky et al., *Man-Machine Integration*, 23-29.

87 Speirs, "Lexicography and Corpus-Tagging," 142-46.

88 But see Gneuss, "The Origin of Standard Old English"; and Gretsch, "Winchester Vocabulary and Standard Old English," 69-83.

89 Stokes and Pierazzo, "Encoding the Language of Landscape," 226-28.

the number of possibilities multiplies enormously, and every one of these possibilities must be anticipated if an automatic system is to be effective. Even if this is done, there remains the problem of homonyms and resulting ambiguities — a problem in all languages, but one which is particularly knotty when combined with potentially wide variations in spelling.[90] For example, Fred Robinson listed headwords which are homonyms of very high-frequency forms: *ac* "but" but also "oak"; *æt* "at" but also "ate" or "food"; *for* "before" but also "journey," "pig," or "went"; *is* "is" but also "ice"; *ofer* "over" but also "seashore."[91] Given that these forms alone occur nearly forty-five thousand times in the *Corpus*, disambiguating just these five headwords — without even considering the problem of variant spellings — would be an arduous task, indeed. The difficulties described here are not new, nor are they unique to Old English, and various computer-aided methods have been developed to ease this process,[92] but the labour required is still very substantial.

Conclusions

A number of broad conclusions and lessons for digital projects can be drawn from the *Dictionary of Old English* and its history. Some of these have already been addressed, such as the importance of releasing both the product itself and the data which lies beneath it to enable unanticipated uses and re-uses of the material. Data which complies to standards also facilitates this re-use and improves its own longevity. Longevity is further assisted by developing software in as high-level and cross-platform a manner as possible: writing the concordancing software in Fortran instead of assembly language extended the project's longevity; in contrast, the "DOESearch" software on the CD-ROM version of the *Dictionary* operates only on Microsoft Windows and almost by definition has a limited lifespan since it depends on a particular operating system which will inevitably change. However, this difficulty is ameliorated by the separation of content and presentation, which is another important lesson. As discussed above, the contents of the *Corpus* and *Dictionary* are now distinct from the presentation: one can access the underlying data without having to use the graphical interface that the project provides. This not only enables the sort of re-use that has already

90 Stokes and Pierazzo, "Encoding the Language of Landscape," 224-25.

91 Fred Robinson, review of *A Microfiche Concordance to Old English*, 134.

92 Venezky, "Computational Aids to Dictionary Compilation," 316-19; Venezky et al., *Man-Machine Integration*, 23-29; Merkin, Busharia, and Meir, "The Historical Dictionary of the Hebrew Language"; Stokes and Pierazzo, "Encoding the Language of Landscape," 226-34.

been discussed, but it also means that the data does not die with the interface: even when the "DOESearch" facility no longer works on future computers, for example, the underlying data will still be accessible. Related to this is a further lesson, namely, that one cannot predict all the uses that other scholars will find for one's digital resource and that as much flexibility and access to the underlying data as possible should be provided. Indeed, in the case of the project to date, the *Corpus* data has arguably proven to be the most important product. Just as one cannot predict how one's resource will be used, so one cannot predict how computing technology will evolve in future, as is vividly demonstrated by the discussion above. Nevertheless, conforming to best practices such as standards-compliance and careful planning, even at the expense of short-term productivity, has proven critical to meeting these challenges over the long term. This is perhaps even more significant today, when the demands of funding bodies, research assessment, and appointment committees privilege short-term projects with rapid results, but the Dictionary of Old English Project has clearly shown the importance of careful planning from the very start. The decade from the initial meeting to the publication of the *Concordance* would be hard to justify today, and even more so the fifteen years to the appearance of the *Dictionary*'s first fascicle, yet it was precisely this care which resulted in a valuable and sustainable resource, a lesson which should not be forgotten. These lessons are all frequently mentioned and recognized in principle; unfortunately, they are still not always followed in practice.[93] The final lesson is perhaps less frequently mentioned, namely, the constant emphasis on placing people before machines, and that scholars in the humanities and in the digital domain worked together from the very start and seemed to genuinely understand each other's needs and limitations. The subtitle of Venezky's article of 1988 emphasizes this by specifying "computer design for a *scholar*'s dictionary" (emphasis mine); and Healey wrote in some detail on computing in her article of 1985 despite her being a scholar of the humanities.[94] Much earlier than this, Peter Clemoes reminded us that any use of a manuscript, including (and perhaps especially) "conversion into computer material," is an act of interpretation; by implication even the most diplomatic transcript cannot be objective or final, a point

93 Besser, "The Past, Present, and Future of Digital Libraries," 564-73; Smith, "Preservation"; and O'Donnell, "Disciplinary Impact and Technological Obsolescence" and "The Doomsday Machine."

94 Venezky, "Unseen Users, Unknown Systems," and Healey, "The Dictionary of Old English and the Final Design of its Computer System." See also Venezky, "Computational Aids to Dictionary Compilation," 309, and Venezky et al., *Man-Machine Integration*, 134-35 and 139.

that is often forgotten today.[95] This mutual understanding of computing and the humanities has become a familiar requirement with today's large projects in Digital Humanities but is not always found in practice.[96]

If anything, the problem with the *Corpus* is its enormous success: as it is used more and as its value is recognized, so scholars try to do more and more with it, making more demands of the project members and sometimes using the resources in ways for which it is not suited. Nevertheless, the careful innovation and perceptive forward-thinking which members of the project have consistently shown is something that we could all hope to emulate.[97]

King's College London

95 Clemoes, speaking on "The Nature of Manuscript Evidence," in Cameron, Frank, and Leyerle, eds. *Computers and Old English Concordances*, 88. For similar views, see Robinson and Solopova, "Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue," 19; Peter Robinson, "What Text Really is Not"; and Pierazzo, "The Limits of Representation," forthcoming; and compare the claims of objectivity expressed in *Editors' Handbook*, §6.i.a12, among others.

96 Pierazzo, "Editorial Teamwork."

97 I wish to thank Toni Healey and Michael Drout for their comments and corrections to parts of this paper. Any errors which remain are, of course, mine alone. I also thank the Leverhulme Trust and the Isaac Newton Trust for their financial support which enabled this research.

Bibliography

Bailey, Richard W. "Research Dictionaries." *American Speech* 44, no. 3 (1969): 166-72.

———, James W. Downer, and Jay L. Robinson, with Patricia V. Lehman. *Michigan Early Modern English Materials*. Ann Arbor, Mich.: Xerox Univ. Microfilms in cooperation with Univ. of Michigan Press, 1975.

———, et al. *Michigan Early Modern English Materials*. Ann Arbor, Mich.: Univ. of Michigan, 2006. <http://quod.lib.umich.edu/m/mememl/index.html>

Besser, Howard. "The Past, Present, and Future of Digital Libraries." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 557-75. Oxford: Blackwell, 2004.

Bratley, Paul, and Serge Lusignan. "Information Processing in Dictionary Making: Some Technical Guidelines." *Computers and the Humanities* 10, no. 3 (1976): 133-43. <http://dx.doi.org/10.1007/BF02426299>

Burton, D[olores] M. "Automated Concordances and Word Indexes: The Fifties." *Computers and the Humanities* 15, no.1 (1981): 1-14. <http://dx.doi.org/10.1007/BF02404370>

———. "Automated Concordances and Word Indexes: The Early Sixties and the Early Centers." *Computers and the Humanities* 15, no. 2 (1981): 83-100. <http://dx.doi.org/10.1007/BF02404202>

———. "Automated Concordances and Word Indexes: The Process, the Programs, and the Products." *Computers and the Humanities* 15, no. 3 (1981): 139-54. <http://dx.doi.org/10.1007/BF02404180>

Busa, R[oberto]. "The Annals of Humanities Computing: The Index Thomisticus." *Computers and the Humanities* 14, no. 2 (1980): 83-90. <http://dx.doi.org/10.1007/BF02403798>

———, ed. *Index Thomisticus: Sancti Thomae Aquinatis operum omnium indices et concordantiae in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur*. Stuttgart-Bad Cannstatt: Frommann-Holzboog, 1974-1980.

———. *Sancti Thomae Aquinatis Hymnorum ritualium varia specimina concordantiarum: primo saggio di indici di parole automaticamente composti e stampati da macchine IBM a schede perforate* (A first example of word index automatically compiled and printed by IBM punched card machines). Milano: Fratelli Bocca, 1951.

———. *Thomae Aquinatis Opera omnia: cum hypertextibus in CD-ROM*. Milan: Editoria Elettronica Editel, 1992.

———, Eduardo Bernot, and Enrique Alarcón. *Corpus Thomisticum: Index Thomisticus, Web Edition*. 2005. <http://www.corpusthomisticum.org/it/>

Cailliau, Robert. *A Little History of the World Wide Web*. World Wide Web Consortium, 2006. <http://www.w3.org/History.html>

Cameron, Angus. "A List of Old English Texts." In *A Plan for the Dictionary of Old English*, edited by Roberta Frank and Angus Cameron, 25-306. Toronto: Univ. of Toronto Press, in association with the Centre for Medieval Studies, Univ. of Toronto, 1973.

———, Ashley Crandell Amos, Sharon Butler, Antonette diPaolo Healey. *The Dictionary of Old English Corpus in Electronic Form.* Toronto: Univ. of Toronto Press, in association with the Centre for Medieval Studies, Univ. of Toronto, 1981. (Distributed on magnetic tape.)

———, Roberta Frank, and John Leyerle, eds. *Computers and Old English Concordances*. Toronto: Univ. of Toronto Press, 1970.

———, Allison Kingsmill, and Ashley Crandell Amos. *Old English Word Studies: A Preliminary Author and Word Index*. Toronto Old English Series 8. Toronto: Univ. of Toronto Press, in association with the Centre for Medieval Studies, Univ. of Toronto, 1983.

de Schryver, Gilles-Maurice. "Lexicographers' Dreams in the Electronic-Dictionary Age." *International Journal of Lexicography* 16, no. 2 (2003): 143-99. <http://tshwanedje.com/publications/dreams.pdf>

Dictionary of American Regional English. [n.d.] <http://dare.wisc.edu/>

Drout, Michael [D. C.]. "Lexomics at Kalamazoo." *Wormtalk and Slugspeak* [2009]. <http://wormtalk.blogspot.com/2009/05/lexomics-at-kalamazoo-do-you-like-to.html>

———, Michael Kahn, Mark D. LeBlanc, Amos Jones, Neil Kathok, and Christina Nelson. "Lexomics for Anglo-Saxon Literature." *Old English Newsletter* 42, no. 1 (2010). <http://www.oenewsletter.org/OEN/pdf/drout42_1.pdf>

———, Michael Kahn, Mark LeBlanc, and Christina Nelson. "Lexomics for Anglo-Saxon Literature." Paper presented at the International Society of Anglo-Saxonists Conference (ISAS), St. John's, Newfoundland, 28 July 2009.

Editors' Handbook – Mini Manual. Skaldic Poetry of the Scandinavian Middle Ages, 2007. <http://skaldic.arts.usyd.edu.au/docs/minimanual.pdf>

Gneuss, Helmut. "Guide to the Editing and Preparation of Texts for the *Dictionary of Old English*." In *A Plan for the Dictionary of Old English*, edited by Roberta Frank and Angus Cameron, 9-24. Toronto: Univ. of Toronto Press, in association with the Centre for Medieval Studies, Univ. of Toronto, 1973.

———. "The Origin of Standard Old English and Æthelwold's School at Winchester." *Anglo-Saxon England* 1 (1972): 63-83.

Godden, Malcolm, ed. *Ælfric's Catholic Homilies: The Second Series – Text*. EETS s.s. 5. London: Oxford Univ. Press, 1979.

Gretsch, Mechthild. "Winchester Vocabulary and Standard Old English: The Vernacular in Late Anglo-Saxon England." *Bulletin of the John Rylands University Library of Manchester* 83, no. 1 (Spring 2001): 41-87.

Healey, Antonette diPaolo. "The Dictionary of Old English and the Final Design of its Computer System." *Computers and the Humanities* 19, no. 4 (1985): 245-49.

———. "The Dictionary of Old English: From Manuscripts to Megabytes." *Dictionaries* 23 (2002): 156-79.

———. "The *Dictionary of Old English*: The Next Generation(s)." In *From Orthography to Pedagogy: Essays in Honor of Richard L. Venezky*, edited by Tom Trabasso, 289-307. Mahwah, N. J.: Lawrence Erlbaum, 2005.

———. "Wood-Gatherers and Cottage-Builders: Old Words and New Ways at the Dictionary of Old English." In *Tracing the Trail of Time: Proceedings from the Second Diachronic Corpora Workshop, New College, University of Toronto, Toronto, May 1995*, edited by Raymond Hickey, Merja Kytö, Ian Lancashire, and Matti Rissanen, 33-46. Language and Computers 18. Amsterdam & Atlanta, Ga.: Rodopi, 1997.

——— et al. *The Dictionary of Old English Corpus in Electronic Form*. University of Michigan Humanities Text Initiative, 2009. [CD-ROM and Web versions]

——— et al. "Documentation on SGML, XML, and HTML Corpus" 2 and 3. 2004. [A PDF document included in the 2004 *Corpus* on CD-ROM.]

Hockey, Susan. "The History of Humanities Computing." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 3-19. Oxford: Blackwell, 2004.

Jenkyns, Joy. "The Toronto *Dictionary of Old English* Resources: A User's View." *Review of English Studies* 42 (1991): 380-416.

Kemble, John Mitchell, ed. *Codex diplomaticus aevi Saxonici*. London: English Historical Society, 1839-1848.

Ker, N. R. *Catalogue of Manuscripts Containing Anglo-Saxon*. Oxford: Clarendon Press, 1957.

LangScape: The Language of Landscape: Reading the Anglo-Saxon Countryside. 2008. <http://www.langscape.org.uk>

LeBlanc, Mark D., et al. *Wheaton College Lexomics Group*. Norton, Mass.: 2009. <http://lexomics.wheatoncollege.edu/>

Leyerle, John. "'The Dictionary of Old English': A Progress Report." *Computers and the Humanities* 5, no. 5 (1971): 279-83. <http://dx.doi.org/10.1007/BF02402209>

Ma'agarim: A Database, Second Century B.C.E. - First Half of the Eleventh Century C.E. Jerusalem: The Academy of the Hebrew Language, 2001.

Ma'agarim: Online Historical Dictionary of Hebrew from the 2nd c. BCE - 11th c. CE. Jerusalem: The Academy of the Hebrew Language, 2009. <http://hebrew-treasures.huji.ac.il/>

Merkin, R., Z. Busharia, and E. Meir. "The Historical Dictionary of the Hebrew Language." *Literary and Linguistic Computing* 4 (1989): 271-73. <http://dx.doi.org/10.1093/llc/4.4.271>

Microsoft Corporation. *Windows History: Windows Desktop Products History*. 2006. <http://www.microsoft.com/windows/WinHistoryDesktop.mspx>

Miller, Thomas, ed. *The Old English Version of Bede's Ecclesiastical History of the English People*. EETS o.s. 95, 96, 110, 111. London: Trübner, 1890-1898.

Mishor, Mordechay. "The Philological Treatment of Ancient Texts: The Experience of the Historical Hebrew Dictionary Project." *International Journal of Lexicography* 15, no. 2 (2002): 132-38. <http://dx.doi.org/10.1093/ijl/15.2.132>

Neuhaus, H. Joachim. "Design Options for a Lexical Database of Old English." In *Problems of Old English Lexicography: Studies in Memory of Angus Cameron*, edited by Alfred Bammesberger, 197-209. Eichstätter Beiträge 15. Regensburg: F. Pustet, 1985.

O'Donnell, Daniel Paul. *Cædmon's Hymn: A Multimedia Study, Archive and Edition*. Woodbridge: D. S. Brewer, 2005.

———. "Disciplinary Impact and Technological Obsolescence in Digital Medieval Studies." In *A Companion to Digital Literary Studies*, edited by Ray Siemens and Susan Schreibman, 65-81. Oxford: Blackwell, 2007.

———. "The Doomsday Machine, or, 'If you build it, will they still come ten years from now?': What Medievalists working in digital media can do to ensure the longevity of their research." *The Heroic Age* 7 (2004). <http://www.mun.ca/mst/heroicage/issues/7/ecolumn.html>

Pierazzo, Elena. "Editorial Teamwork in a Digital Environment: The Edition of the Correspondence of Giacomo Puccini." *Jahrbuch für Computerphilologie* 10 (2008): 91-109. <http://computerphilologie.tu-darmstadt.de/jg08/pierazzo.html>

———. "The Limits of Representation: A Rationale of Digital Documentary Editions." Forthcoming.

Price-Wilkin, John. "A Campus-Wide Textual Analysis Server: Projects, Prospects, and Problems." In *Proceedings of the 8th Annual Conference of the Centre for the New Oxford English Dictionary and Text Research*. Waterloo, Ont.: Univ. of Waterloo Centre for the New OED, 1992. <http://jpw.umdl.umich.edu/pubs/waterloo.html>

Rennie, Susan, et al., eds. *Dictionary of the Scots Language*. Dundee: Univ. of Dundee, 2004. <http://www.dsl.ac.uk/>

Robinson, Fred C. Review of *A Microfiche Concordance to Old English*, compiled by Antonette diPaolo Healey and Richard L. Venezky. *Speculum* 57, no. 1 (1982): 133-35.

Robinson, Peter. "What Text Really is Not, and Why Editors Have to Learn to Swim." *Literary and Linguistic Computing* 24, no. 1 (2009): 41-52. <http://dx.doi.org/10.1093/llc/fqn030>

——— and Elizabeth Solopova. "Guidelines for Transcription of the Manuscripts of the Wife of Bath's Prologue." In *The Canterbury Tales Project: Occasional Papers*, edited by Norman Blake and Peter Robinson, 1:19-52. Oxford: Oxford Univ. Computing Services, Office for Humanities Communication, 1993. <http://www.canterburytalesproject.org/pubs/op1-transguide.pdf>

Scragg, Donald, et al. *ManCASS C11 Database Project*. Manchester: Manchester Centre for Anglo-Saxon Studies, 2005. <http://www.arts.manchester.ac.uk/mancass/C11database/>

Smith, Abbey. "Preservation." In *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth, 576-91. Oxford: Blackwell, 2004.

Speirs, Nancy. "Lexicography and Corpus-Tagging: Enhancing *The Dictionary of Old English Corpus in Electronic Form*." In *Tracing the Trail of Time: Proceedings from the Second Diachronic Corpora Workshop, New College, University of Toronto, Toronto, May 1995*, edited by Raymond Hickey, Merja Kytö, Ian Lancashire, and Matti Rissanen, 137-49. Language and Computers 18. Amsterdam & Atlanta, Ga.: Rodopi, 1997.

Sperry Rand. *Univac 1108 II*. Sales brochure for Sperry Rand Corporation. 1965. <http://archive.computerhistory.org/resources/text/Remington_Rand/SperryRand.UNIVAC1108II.1965.102646105.pdf>

Stokes, Peter A. "Computer-Aided Palaeography, Present and Future." In *Kodikologie und Paläographie im digitalen Zeitalter – Codicology and Palaeography in the Digital Age*, edited by Malte Rehbein, Patrick Sahle, and Torsten Schaßan, 309-38. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt: Books on Demand, 2009.

———. Review of *Cædmon's Hymn: A Multimedia Study, Edition and Archive* [*sic*], by Daniel Paul O'Donnell. *Digital Medievalist* 5 (2009). <http://www.digitalmedievalist.org/journal/5/stokes/>

——— and Elena Pierazzo. "Encoding the Language of Landscape: XML and Databases at the Service of Anglo-Saxon Lexicography." In *Perspectives on Lexicography in Italy and Europe*, edited by Silvia Bruti, Roberta Cella, and Marina Foschi Albert, 203-38. Newcastle upon Tyne: Cambridge Scholars, 2009.

Taylor, Ann, et al., eds. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*. Oxford: Oxford Text Archive, 2003. <http://ota.oucs.ox.ac.uk/headers/2462.xml>

TEI Consortium, The. *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Edited by C. M. Sperberg-McQueen and Lou Burnard. Ann Arbor, Mich.: Univ. of Michigan, 1999. <http://quod.lib.umich.edu/t/tei/>

TEI Consortium, The. *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Edited by C. M. Sperberg-McQueen and Lou Burnard. Oxford: Humanities Computing Unit, Univ. of Oxford, 2004. <http://www.tei-c.org/release/doc/tei-p4-doc/html/>

TEI Consortium, The. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Edited by Lou Burnard and Syd Bauman. Oxford: Humanities Computing Unit, Univ. of Oxford, 2009. <http://www.tei-c.org/Guidelines/P5/>

Venezky, Richard L. "Computational Aids to Dictionary Compilation." In *A Plan for the Dictionary of Old English*, edited by Roberta Frank and Angus Cameron, 307-27. Toronto: Univ. of Toronto Press, in association with the Centre for Medieval Studies, Univ. of Toronto, 1973.

———. "Unseen Users, Unknown Systems: Computer Design for a Scholar's Dictionary." *Computers and the Humanities* 22, no. 4 (1988): 285-91.

——— and Antonette diPaolo Healey. *A Microfiche Concordance to Old English*. Publications of the Dictionary of Old English 1. Toronto: Pontifical Institute of Mediaeval Studies, 1980.

——— et al. *Man-Machine Integration in a Lexical Processing System*. Computer Sciences Technical Report #280. Madison: Univ. of Wisconsin, Computer Sciences Department, 1976. <http://ftp.cs.wisc.edu/pub/techreports/1976/TR280.pdf>

Walker, John. *Typical UNIVAC 1108 Prices: 1968*. 1996. <http://www.fourmilab.ch/documents/univac/config1108.html>