Renaissance and Reformation
Renaissance et Réforme

Word-entries and Big Data in Lexicons of Early Modern English

Ian Lancashire

Volume 37, numéro 4, automne 2014

In Celebration of the Fiftieth Anniversary (II)
En célébration du cinquantenaire 1964-2014 (II)

URI : https://id.erudit.org/iderudit/1090643ar
DOI : https://doi.org/10.33137/rr.v37i4.22648

Aller au sommaire du numéro

Éditeur(s)

Iter Press

ISSN

0034-429X (imprimé)
2293-7374 (numérique)

Découvrir la revue

Citer cet article

Lancashire, I. (2014). Word-entries and Big Data in Lexicons of Early Modern English. *Renaissance and Reformation / Renaissance et Réforme*, *37*(4), 215–233. https://doi.org/10.33137/rr.v37i4.22648

Résumé de l'article

Cette brève histoire des trente ans du Lexicons of Early Modern English, une base de données en ligne de glossaires et de dictionnaires de l'époque, commence en 1986 dans le laboratoire du Centre for Computing and the Humanities, au quatorzième étage de la bibliothèque Robarts de l'Université de Toronto. Cette base de données a été publiée gratuitement en ligne premièrement en 1996, sous le titre Early Modern English Dictionnaires Database. Dix ans plus tard, elle était publiée sous le sigle LEME, à partir du septième étage de la même bibliothèque Robarts, grâce au soutien du TAPoR (Text Analysis Portal for Research), de la bibliothèque et des presses de l'Université de Toronto. Aucune autre langue vivante ne dispose d'une telle ressource. La principale raison expliquant l'émergence, la survie et la croissance du LEME est que les lexicographes qui font l'objet du LEME comprenaient leur langue très différemment que nous la concevons depuis deux siècles, et ce nonobstant plusieurs de nos avantages.

Word-entries and Big Data in *Lexicons of Early Modern English*

IAN LANCASHIRE
Robarts Library, University of Toronto

This brief thirty-year history of Lexicons of Early Modern English, *an online database of glossaries and dictionaries of the period, begins in a fourteenth-floor Robarts Library lab of the Centre for Computing and the Humanities at the University of Toronto in 1986. It was first published freely online in 1996 as the* Early Modern English Dictionaries Database. *Ten years later, in a seventh-floor lab also in the Robarts Library, it came out as* LEME, *thanks to support from TAPoR (Text Analysis Portal for Research) and the University of Toronto Press and Library. No other modern language has such a resource. The most important reason for the emergence, survival, and growth of* LEME *is that its contemporary lexicographers understood their language differently from how we, our many advantages notwithstanding, have conceived it over the past two centuries.*

Cette brève histoire des trente ans du Lexicons of Early Modern English, *une base de données en ligne de glossaires et de dictionnaires de l'époque, commence en 1986 dans le laboratoire du Centre for Computing and the Humanities, au quatorzième étage de la bibliothèque Robarts de l'Université de Toronto. Cette base de données a été publiée gratuitement en ligne premièrement en 1996, sous le titre* Early Modern English Dictionnaires Database. *Dix ans plus tard, elle était publiée sous le sigle* LEME, *à partir du septième étage de la même bibliothèque Robarts, grâce au soutien du TAPoR (Text Analysis Portal for Research), de la bibliothèque et des presses de l'Université de Toronto. Aucune autre langue vivante ne dispose d'une telle ressource. La principale raison expliquant l'émergence, la survie et la croissance du* LEME *est que les lexicographes qui font l'objet du* LEME *comprenaient leur langue très différemment que nous la concevons depuis deux siècles, et ce nonobstant plusieurs de nos avantages.*

Lexicons of Early Modern English (*LEME*, 2006–) is a made-in-Canada online historical database of 713,000 word-entries from 203 dictionaries and glossaries printed or written in England from 1475 to 1755.[1] Its growth, from a

1. Ian Lancashire, ed., *The Lexicons of Early Modern English* [*LEME*] (Toronto: University of Toronto Press and the University of Toronto Library, 2006–), http://leme.library.utoronto.ca. In mid-2014, *LEME* welcomed Carol Percy as associate editor, eighteenth century, and extended its coverage from 1702 to 1755. *LEME* is grateful for assistance by the Social Sciences and Humanities Research Council of Canada (SSHRC), the Canada Foundation for Innovation (CFI), the Text Analysis Portal for Research (TAPoR), the Early English Books Online/Text Creation Partnership (EEBO/TCP), the Internet Archive, and the University of Toronto. An earlier version of this paper, "Why Lexicons of Early Modern English?," was

single dictionary text begun in 1986 to a tool for interrogating big data today, illustrates how emerging technology and individual research can make a sizable digital humanities resource. First published freely online as the *Early Modern English Dictionaries Database* (*EMEDD*) in 1996,[2] *LEME* is now hosted by the University of Toronto Library and published by the University of Toronto Press.[3] Its corpus reveals contemporary English vocabulary, dominant word-senses, typical word-spellings, and dates of usage, particularly in bilingual dictionaries for Latin, French, Italian, Spanish, Welsh, and Old English. Because many of these lexicons alphabetized entries by foreign-language headwords, the only way before *EMEDD/LEME* to locate English words employed as synonyms or translations in a word-entry's post-lemmatic position was to read a lexicon manually from start to finish. *LEME* also includes a bibliography of about 1,300 lexical sources for the period.

LEME pre-history

LEME is a late-life child of humanities computing in the late 1970s, when a group of University of Toronto researchers requested that the university's Computing Services hire a programmer to support concordancing of large texts

---

given at the Nineteenth Biennial Meeting of the Dictionary Society of North America held at the University of Georgia in Athens, Georgia, on 24 May 2013. For comments on *LEME*, see Ray Siemens and Gary Shawver, "Introduction: A Volume Celebrating and Recognizing Ian Lancashire," *Digital Studies* 1.1 (2009; *LEME* is "helping to redefine our understanding of the shape of our language"), http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/163; David Vancil, "Seven North American Dictionary Collections," *Dictionaries: Journal of the Dictionary Society of North America* 32 (2011): 111–28 ("one of the best-designed research websites I have encountered"); Gabriele Stein, *Sir Thomas Elyot as Lexicographer* (Oxford: Clarendon Press, 2014): 17 ("superb and unparalleled"); Hardy M. Cook, "New Entries in Lexicons of Early Modern English," *The Shakespeare Conference: SHK* 22.0302 (16 November 2011; "an invaluable tool for annotating"); Oxford English Dictionary, "Consultants, advisors, and contributors," *Rewriting the* OED (2013; *EMEDD* and *LEME* "have played a significant part in the work"); and Patrick Hanks, "Lexicography, Printing Technology, and the Spread of Renaissance Culture" (2010; "magnificent"), http://www.patrickhanks.com/uploads/5/1/4/9/5149363/renaissancelexicography.pdf. I am grateful to Peter Gilliver for his most helpful and genial advice on this paper, and to my three anonymous reviewers for their insightful suggestions.

2. Ian Lancashire, ed., *The Early Modern English Dictionaries Database* [*EMEDD*] (University of Toronto: Computers in the Humanities and Social Sciences, 1996–99) (site discontinued).

3. *LEME* is free for occasional queries, but for sustained research it should be licensed by an institution or an individual. Fees support web technology and enable *LEME* to add new lexical works every year.

on its IBM mainframe.[4] John Bradley (now at King's College London), who had developed a multilingual concording engine at the University of Waterloo, joined Toronto's Computing Services and implemented a concordancer named COGS ("COncordance-Generating System"). Faculty researchers who were interested in it included John Hurd in Divinity, Jack Stevenson in Philosophy, Barron Brainerd in Statistics, Russon Wooldridge in French, and myself. Joseph Raben, editor of the journal *Computers in the Humanities*, had encouraged hundreds of scholars to use database technology in controlling information, such as a list of medieval English lexical texts (Huntsman 1978), Early Modern English play titles (Berger and Struminger 1985), and an index of historical records of play performance (Lancashire 1978).[5] However, products like Oracle would later serve, better than COGS, the need to control bibliographical information.[6] COGS produced concordances of full texts written in European vernacular languages.

Concording technology goes back to the Key Word In Context (KWIC) index developed in the late 1950s (Williams 2010) and the (printed) Cornell concordances that had applied it effectively in indexing major authors (Parrish 1959).[7] To analyze the great Renaissance French dictionaries by Estienne and Nicot, Russon Wooldridge needed what we called a textbase,[8] and Bradley's concordance generator was ideal for that purpose. It also had potential in helping

4. The documents recording this and other historical information in this article are to be donated to University Archives, University of Toronto.

5. Jeffrey F. Huntsman, "Computers and Medieval English Lexicography," *Computers and the Humanities* 12 (1978): 53–60; Thomas L. Berger and Leny Struminger, "Panel Discussion on the Creation of a Database for the Drama of the English Renaissance," *The International Conference on Data Bases in the Humanities and Social Sciences*, 1983, ed. Robert F. Allen (Osprey, Florida: Paradigm Press, 1985), 160–64; and Ian Lancashire, "Records of Early English Drama and the Computer," *Computers and the Humanities* 12 (1978): 183–88.

6. I needed an indexing program that Willard McCarty, then at Records of Early English Drama, wrote in FORTRAN for building the numbered indexes in my *Dramatic Texts and Records of Britain: A Chronological Topography* (Toronto: University of Toronto Press, 1984).

7. Robert V. Williams, "Hans Peter Luhn and Herbert M. Ohlman: Their Roles in the Origins of Keyword-in-Context/Permutation Automatic Indexing," *Journal of the American Society for Information Science and Technology* 61.4 (2010): 835–49; and Stephen Maxfield Parrish, *A Concordance to the Poems of Matthew Arnold* (Ithaca, NY: Cornell University Press, 1959).

8. Terence R. Woolridge, "The Estienne-Nicot Project: An Inventory of the French Lexicon of the Principal Dictionaries of the Sixteenth Century," in *The International Conference on Data Bases,* ed. Allen, 43–47.

students appreciate the repetition and variation of vocabulary in literary texts. I then began teaching students how to use computers (Lancashire 1983), and worked with Bradley's assistant, Lidio Presutti, to produce *Microcomputer Text Analysis System* (*MTAS*, a text-searching and analysis program for the IBM PC) in 1984 that applied concording principles to small texts.[9] An anticipation of how useful this might be in the not-too-distant future of technology led the University of Toronto in 1985 to found the Centre for Computing in the Humanities (CCH) under my direction.[10] CCH benefitted from a cooperative with IBM Canada Ltd., signed in 1986, one university deliverable for which was Toronto's microcomputer-based version of Bradley's COGS, an interactive concordancer we called *Text Analysis Computing Tools* (*TACT*).[11] This was freely released in 1989 when CCH hosted the first joint International Conference on Computing in the Humanities (ICCH) and the Association for Literary and Linguistic Computing (ALLC) at Toronto. *TACT* followed the groundbreaking *Oxford Concordance Program* and *Wordcruncher*, a fine commercial product.

The first textbase I developed for use with *MTAS* and *TACT* was a collection of twenty-six early Tudor moral plays and interludes, and John Palsgrave's English-French dictionary, *Lesclarcissement* (1530). I had begun text-entry for this lexicon—the seed for *LEME*—in early 1986 and finished in 1989.[12] My first book, *Two Tudor Interludes*, had been a Revels Plays edition (1980), and I thought that Palsgrave offered a useful tool to interpret the vocabulary of early play-texts. My interest in historical lexicography came from my association with Wooldridge, as well as with C. C. (Kelly) Gotlieb, the founder of computer science in Canada, who introduced me to the University of Waterloo *Oxford English*

9. Ian Lancashire, *Computer Applications in Literary Studies: A Userbook for Students at Toronto* (Toronto: Department of English, 1983). *Micro TextAnalysis System* (*MTAS*), programmed by Lidio Presutti in Turbo Pascal for the IBM PC, came out in 1985, and version 2.0 followed in March 1988. Program and source code were published on the Internet in February 1993.

10. CCH became Computing in the Humanities and the Social Sciences (CHASS) in 1996.

11. Ian Lancashire, John Bradley, Willard McCarty, Michael Stairs, and T. R. Wooldridge, *Using* TACT *with Electronic Texts: A Guide to* Text-Analysis Computing Tools, *Version 2.1 for MS-DOS and PC DOS* (New York: Modern Language Association of America, 1996); online at http://www.mla.org/store/CID7/PID236. Bradley discusses COGS and TACT in his "What the Developer Saw: An Outsider's View of Annotation, Interpretation and Scholarship," *Digital Studies* 1.1 (2009), http://www.digitalstudies.org/ojs/index.php/digital_studies/article/view/143/202.

12. For this I received a grant of $1,785 from the Humanities and Social Sciences Committee of the University of Toronto on 30 January 1986.

Dictionary (*OED*) project, where I gave a paper in 1987.[13] By that summer, I visited the University of Otago in Dunedin, New Zealand, to discuss collaboration with Alistair Fox and Greg Waite, who were developing an Early Tudor poetry textbase.[14] I mailed my e-texts to Fox in March 1989 and promised to convene a meeting at the ALLC/ICCH89 in Toronto to discuss a Renaissance textbase project.

Roy Flannagan (the editor of Milton's poetry) and I wrote the draft proposal for a Renaissance textbase, and we and Lou Burnard (Oxford), Thomas Corns (Bangor), and David Richardson (Cleveland State; editor of the Spenser encyclopedia) co-presented it at the ALLC/ICCH89 conference on 6 June.[15] By December 1989 our proposal covered selected major authors and "texts representative of the language and thought of the period" such as dictionaries by Randle Cotgrave (French), Thomas Cooper (Latin), John Florio (Italian), William Salesbury (Welsh), and John Minsheu (Spanish). The textbase was to use TEI SGML markup, to be archived at the Oxford Text Archive (which I did in 1990), and to jump-start with a National Endowment for the Humanities (NEH) application by Flannagan, followed by applications to the Social Sciences and Humanities Research Council (SSHRC) by myself, and to the British Academy by Corns. We were assembling a larger advisory group (to which Canadians A. C. Hamilton and Frank Tompa belonged) when we learned that NEH had declined Flannagan's request to fund the group's modest travel support because we could not promise "sustained institutional funding."

David Richardson then took up the initiative and applied to NEH on our behalf for an emergency planning grant to prepare a revised proposal, now renamed—largely because of the leverage that dictionaries offered—the Renaissance Knowledge Base (Siemens and others, 2011).[16] In late December

13. Ian Lancashire, "Using a Textbase for English Language Research," in *The Uses of Large Text Databases,* Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary, 9–10 November 1987 (Waterloo, ON: University of Waterloo Centre for the New Oxford English Dictionary, 1987), 51–64.

14. Greg Waite, ed., *The Textbase of Early Tudor English* (University of Otago, NZ, 1995), http://www.otago.ac.nz/englishlinguistics/english/Tudor%20Pages%202013/home.html.

15. The documents recording this information are to be donated to University Archives, University of Toronto.

16. Ray Siemens, Mike Elkink, Alastair McColl, Karin Armstrong, James Dixon, Angelsea Saby, Breet D. Hirsch and Cara Leitch, with Martin Holmes, Eric Haswell, Chris Gaudet, Paul Girn, Michael Joyce,

1990 we learned that NEH funding was again refused, and for the same reason. We next convened an open meeting of the Renaissance Knowledge Base on 18 March 1991, at the third joint conference of the Association for Computing and the Humanities (ACH) and the ALLC in Tempe, Arizona.[17] This session featured Eric Calaluca (marketing vice-president of Chadwyck-Healey), discussing "large-scale plans to transcribe books from the Cambridge Bibliography of English Literature to a storage medium such as CD-ROM." The necessary sustained funding for a digital collection much larger than the Renaissance project, of concern to NEH, would come from industry. Chadwyck-Healey's collection of 1,350 poets from 600 to 1900 was published on 31 October 1991 as *The English Full-text Database*, and later as *Literature Online*. On the eve of the Web, which Tim Berners-Lee had proposed in March 1989 and implemented first in a website on 7 August 1991, we realized why our joint efforts to that date had been unsuccessful.[18] Someone had effectively saved us competing with an industry that had deep pockets.

The completion of the IBM Canada Ltd. cooperative kept me busy, but a dictionaries knowledge base, not part of Chadwyck-Healey's product, remained on my mind. I asked a colleague in California whether anyone trying to develop an Early Modern English dictionaries database would be making a mistake. He replied that the human costs of data entry, proofreading, encoding, and software development were too steep. Most literary texts could be scanned from modern print editions and readily converted with OCR. Early dictionaries, however, resisted quick digitization. They were multilingual, often impossible to OCR, and widely varying in form. Granting agencies would be reluctant to support any such venture in a sustained way because the project would not be primary research but rather tool development. Anyway, he argued, the early dictionaries were frequently erroneous and plagiarized: they paled into insignificance in comparison to the *OED*.

Rachel Gold, and Gerry Watson, and members of the PKP, Iter, TAPoR, and INKE teams, "Prototyping the Renaissance English Knowledgebase (REKn) and Professional Reading Environment (PReE), Past, Present, and Future Concerns: A Digital Humanities Project Narrative," *Digital Studies* 2.2 (2011).

17. Ian Lancashire, "The Toronto Renaissance Dictionaries," in a session on The Renaissance Knowledge Base at the third joint conference of the ACH and ALLC (Tempe, Arizona, 17–21 March 1991).

18. Lou Burnard, our indispensible expert in SGML, was on the Chadwyck-Healey advisory board for this product.

Despite two rejections from NEH, and my colleague's not unrealistic assessment of my plan, I went ahead alone. Fortunately, because of a by-then credible background in early modern English drama and computing technology, SSHRC awarded me a three-year grant of $100,000 in 1990 to do research in content analysis in early modern English texts. I described the kernel of the corpus for content analysis as an English Renaissance Knowledge Base (Lancashire 1992) that comprised literary authors and prose texts such as sermons and dictionaries.[19] On 10 July 1993, I delivered a "Medieval and Renaissance Textbase" that included eight dictionaries input by Gail Richardson (Palsgrave, Randle Cotgrave, and Thomas Thomas), Computer Input Corporation (John Florio), Rosemary Newman (Sir Thomas Elyot), Maria Dumity (William Thomas), and Sharine Leung (Henry Cockeram).[20] Ray Siemens contributed Robert Cawdrey's famous hard-word dictionary (1604), a project he had done for Willard McCarty's graduate course at CCH. The lexical texts were then intended as a resource for interpreting major literary authors.

At the close of this grant, I realized that a dictionaries database was separable from a literary one; even if the digitization of literature had become a business with which I could not compete, language studies and experimentation on literature were still possible. Later in that year Flannagan agreed to serve on the editorial advisory board of my proposed Toronto series, *Renaissance Electronic Texts* (1994–98; *RET*), which I intended for non-lexical works.[21] *RET* issued only the Elizabethan homilies, Shakespeare's sonnets (with Hardy Cook), and Edmund Coote's *The English Schoolmaster* (with my graduate students). Once I published Coote's book, I decided to concentrate on dictionaries. SSHRC

19. Ian Lancashire, "Bilingual Dictionaries in an English Renaissance Knowledge Base," in *Historical Dictionary Databases*, ed. T. R. Wooldridge, CCH Working Papers 2 (Toronto: Centre for Computing in the Humanities, 1992), 69–88; http://www.chass.utoronto.ca/epc/chwp/lancash1/.

20. Ian Lancashire, "A Textbase of Early Modern English Dictionaries 1499–1659" (Georgetown University, Washington, DC, ACH/ALLC International Conference, 16–19 June 1993); and "The Early Modern English Renaissance Dictionaries Corpus," in *English Language Corpora*, ed. J. Aarts, P. de Haan, and N. Oostdijk, *Language and Computers* 10 (Amsterdam and Atlanta: Rodopi, 1993), 11–24.

21. Ian Lancashire, ed., *Renaissance Electronic Texts* (Toronto: University of Toronto Library, 1994–98), http://www.library.utoronto.ca/www/utel/ret/ret.html. See also my "Editing English Renaissance Electronic Texts," in *The Literary Text in the Digital Age*, ed. Richard J. Finneran (Ann Arbor: University of Michigan Press, 1996), 117–43, and "Encoding Renaissance Electronic Texts," in *New Technologies and Renaissance Studies*, ed. William Bowen and Ray Siemens (Iter and the Arizona Center for Medieval and Renaissance Studies, 2009), 243–60.

awarded me a second three-year research grant of $50,000, on 29 March 1994, to create An Early Modern English Dictionary Textbase. Co-hosting, with Wooldridge, a CCH conference on historical dictionaries at Toronto that year gave me increased confidence.[22] By the fall of 1995, I was far enough along in this work to teach a graduate course on Shakespeare's language for which I used my dictionaries as source materials. One of my students was Jennifer Roberts-Smith, who went on to become a *LEME* research assistant. A Computer Science graduate student taking the course, Mark Catt, wrote a program called Patterweb that made online searching of the lexical texts a breeze.[23] On 28 April 1996, as I stepped down from directing CCH after eleven years, I was able to place online some 127,000 word-entries from the dictionaries as the *EMEDD*, freely searchable with Mark Catt's program. In a few years, *EMEDD* had sixteen lexicons and some two thousand researchers registered for its use.

By 1999, I had obtained about $150,000 in research funds for the dictionaries. After SSHRC in 2000 awarded me a final grant for $100,000 to extend the dictionaries corpus, Geoffrey Rockwell at McMaster University proposed, to the Canada Foundation for Innovation (CFI), a project titled Text Analysis Portal for Research (TAPoR). This attracted enthusiastic partners at the University of Toronto, l'Université de Montréal, the University of Alberta, and the University of Victoria; and after several years of negotiation, TAPoR received its multi-million-dollar infrastructural grant.[24] As its Toronto principal investigator, I acquired a very solid institutional backing for *LEME*, conceived as one of the TAPoR deliverables. But the respect that NEH or its assessors evidently had for

22. Ian Lancashire, "An Early Modern English Dictionaries Corpus 1499–1659," in *Early Dictionary Databases*, ed. Ian Lancashire and T. Russon Wooldridge, CCH Working Papers 4 (University of Toronto: Centre for Computing in the Humanities, 1994), http://www.chass.utoronto.ca/epc/chwp/lancash2/; and "The Early Modern English Renaissance Dictionaries Corpus: An Update," in *Corpora across the Centuries*, ed. M. Kyto, Matti Rissanen, and Susan Wright (Amsterdam: Rodopi, 1994), 143–49.

23. Patterweb was a front-end to the Waterloo Pat software, for which see Gaston H. Gonnet, ed., *Examples of PAT applied to the Oxford English Dictionary* (Waterloo, ON: University of Waterloo Centre for the New Oxford English Dictionary, 1987). In 1997 Catt also employed Pat to index the University of Toronto English Library, a collection of e-texts partly derived from the CD-ROM offered with the manual for *TACT* published by MLA. See also Mark Catt, "Renaissance Dictionaries and Shakespeare's Language: A Study of Word-meaning in *Troilus and Cressida*," *Early Modern Literary Studies* Special Issue 1 (1997), 3: 1–46, http://purl.oclc.org/emls/si-01/si-01catt.html.

24. Geoffrey Rockwell, Stefan Sinclair, and K. C. Uszkalo, *TAPoR 2.0* (Edmonton: University of Alberta, 2006–), http://www.tapor.ca.

industry-led infrastructure a decade earlier could be observed in CFI's expectation that TAPoR principals would only be funded if their deliverables had a demonstrable commercial value. I estimated the value of what would become *LEME* by using the Middle English Dictionary as a model (at that time, it was still being licensed), and I sought out the University of Toronto Press as publisher.

A team of University of Toronto Library programmers and digital librarians then developed the software for *LEME*, and the Press managed distribution and licensing fees. We provided free access to the five TAPoR institutions and made casual searches by individuals elsewhere free. *LEME* programmer Marc Plamondon, today directing a digital humanities program at Nipissing University, and designer Sian Meikle, now the head of IT at the University of Toronto Library, created a sturdy digital foundation for *LEME* with SQL and ColdFusion. A laboratory was set up in the IT area of Robarts Library for my work. At release in 2006, *LEME* had 150 lexical texts and half a million word-entries, making it more than twice as big as *EMEDD*. It possessed a fine advisory board and the loyal support of one of the top academic libraries in the world. Yet I had digitized less than 12 percent of the *LEME* bibliography of primary lexical texts; and the University of Toronto Press regarded *LEME* as a journal, that is, as a continuing publication that needed a fresh injection of new content every year.

Justifying *LEME*

After receiving generous research grants and a state-of-the-art institutional infrastructure, to ask whether a project is useful seems disingenuous. However, deactivating *EMEDD*, an entirely free resource, a few years after *LEME* went online, showed that enthusiasm for something free did not always translate into client willingness to pay for an improved version. I had mounted a plausible case for *EMEDD* as a source of some new readings in Shakespeare's work,[25] and as a way of revisiting the making of English in the early modern period.[26] However,

25. Ian Lancashire, "Understanding Shakespeare's *Titus Andronicus* and the EMEDD," in *New Scholarship from Old Renaissance Dictionaries: Applications of the Early Modern English Dictionaries Database*, ed. Ian Lancashire and Michael Best, *Early Modern Literary Studies* (April 1997), http://purl.oclc.org/emls/emlshome.html.

26. During this period I published several general articles on *LEME* development, including "The Lexicons of Early Modern English," *TEXT Technology* 12 (2003) and "Computing the Lexicons of Early Modern

I was faced with a backlash from researchers who expected all texts to be free. On 7 February 2002, *EMEDD* had to be closed down for two weeks and reprogrammed by Tak Ariga of Computing in the Humanities and Social Sciences (CHASS) because a non-Canadian user interested in dance terminology had illegally entered the site and copied parts of the database, notably a major Florio dictionary. It was then released freely online without my permission. I was then faced with several challenges: making the project self-supporting, protecting the data, and justifying why anyone who had the *OED Online* needed *LEME*.

Whether *LEME* is a collection of early lexical word-entries, or a historical dictionary database (Vancil 2011), it is not the *OED Online*. I have shared *LEME* freely with Oxford lexicographers, and it adds to the *OED Online* much Early Modern English lexical material that may improve our understanding of the chronology of vocabulary change.[27] *LEME* also supplies "contemporary comments," the historical information that the *OED* has underrepresented and that C. C. Fries, the first editor of the now-lapsed *Early Modern English Dictionary* at the University of Michigan, stressed the importance of including *in a period dictionary.*[28] The *OED Online* does not usually document misleading statements about words by early lexicographers. *LEME* and the *OED Online* have about the same number of "word-entries," but *LEME* adduces many word-entries with the same lemma, and the *OED Online* has five times

English," in *The Changing Face of Corpus Linguistics*, ed. Antoinette Renouf, Language and Computers: Studies in Practical Linguistics (Amsterdam: Rodopi, 2006), 45–62. I also did some new research on manuscript dictionaries and glossaries, law lexicons, hard words, and the failure of Tudor England to undertake a monolingual English dictionary. See "Lexicography in the Early Modern English Period: The Manuscript Record," in *Historical Lexicography,* ed. J. Coleman and A. MacDermott (Tübingen: Max Niermeyer, 2005), 19–30; "Law and Early Modern English Lexicons," in *HEL-LEX: New Approaches in English Historical Lexis*, ed. Roderick McConchie, Heli Tissari, and Olga Timofeeva (Somerville, MA: Cascadilla Press, 2006), 8–23; "The Two Tongues of Early Modern English," in *Managing Chaos*, ed. Christopher Cain, Studies in the History of English 3 (Berlin: Mouton de Gruyter, 2007), 115–53; and "Why did Tudor England have no Monolingual English Dictionary?" in *Webs of Words: New Studies in Historical Lexicology*, ed. John Considine (Cambridge: Cambridge Scholars Publishing, 2010), 8–23.

27. The *OED Online* cites word-entries from earlier dictionaries only when they uniquely document first occurrences (of words, or of senses). General lexicographical practice today does not acknowledge definitions appearing in earlier dictionaries.

28. Richard W. Bailey, "Charles C. Fries and the *Early Modern English Dictionary*," in *Toward an Understanding of Language: Charles Carpenter Fries in Perspective*, ed. Nancy M. Fries (Amsterdam: Benjamins, 1985), 171–204.

as many quotations (three million), although they cover a much longer time period. The actual overlap between the two is small: under 5 percent of *LEME* word-entries appear in the *OED Online*.

The *OED Online* is informed by a modern theory of semantic development. Oxford lexicographers choose headwords, assign standard spellings to them, select supporting historical quotations, distinguish senses, devise definitions, and research etymologies; and they know much more about Early Modern English than those who spoke and wrote it. The *OED Online* publishes *lexicographical research*. *LEME* word-entries, in contrast, are edited primary historical records: they may be factually wrong about the language at times, but they are arguably more faithful than the late-Victorian, Modern, and New-Century *OED* to the language attitudes of the period because they do not reconstruct early word-meaning according to our present-day understanding, however correct ours may be. Because the *OED Online* has a responsibility to document English words in all periods and regions, it must devise a rigorous method for describing them all across time and space, and it cannot impose one period's language ideas on the entire language history. Anyone who believes that the early modern English regarded language as we have since 1800 is thinking wishfully. Ancient conceptions about language should be taken into account in understanding texts of the period. Consequently, unlike an Oxford lexicographer, I am a servant of Samuel Johnson's drudges[29]—early lexicographers—by entering their texts, proofing them, identifying the language of each word in the text, and, if possible, locating the modern English spelling of headwords and explanatory synonyms, and noting if a headword does not appear in the *OED Online* or antedates the earliest citation there.

Unlike *LEME*, and for good reasons, the *OED Online* excludes several classes of words. Wikipedia, gazetteers, and first-name glossaries explain words that the *OED Online* leaves out. It does not register most place and proper names,[30] encyclopedic information, and a farrago of oddities, questionable headwords, and personal lexical inventions. Neither does the *OED Online*

29. Johnson defines the term "lexicographer" famously as "A writer of dictionaries; a harmless drudge, that busies himself in tracing the original, and detailing the signification of words." See Samuel Johnson, *A Dictionary of the English Language on CD-ROM*, ed. Anne McDermott (Cambridge: Cambridge University Press, 1996).

30. I mean generic proper names such as "Benjamin," not the full names of individuals such as "John Minsheu."

offer a network of foreign-language terms associated with English words that are employed to translate them. *LEME* includes all these, *as those alive in the early modern period presented them*, and also indicates which words were, in their time, considered "hard," and which "easy."[31] It shows, by sheer repetition of entries, which meanings of a word were most common. *LEME* delivers the original image of English in the chaotic form it took when so many different individuals—especially powerful patrons such as Henry VIII and Lord Burghley—contributed their say to what the language would become and before standardization won out.[32] The *OED Online* lexicographers have everything that early modern lexicographers lacked—a large budget, information technology that can search a vast quantity of texts, advanced research training, and three hundred years of scholarship—except for one thing: the in-depth experience of a native speaker alive in the period.

LEME word-entries vary greatly in form and style and can be amusingly anecdotal. For example, Claudius Hollyband's French-English dictionary (1593)[33] explains the word "bougre, he that committed such a fact and sodomite villanie: a buggerer: burne them all." Here the shocking indignation of the glossographer or his compositor bursts through. And lexicographers such as William Camden (1605) describe "Gertrude," the name of Hamlet's mother, incorrectly as meaning "All truth" and "All true, and amiable."[34] Reference sources today translate it properly as "strong spear." Which explanation, the erroneous "all truth" or the accurate "strong spear," best illustrates why Shakespeare kept Gertrude as the name of Hamlet's mother? Should we not be interested in such misinformation? In my opinion, lexicature[35] such as this can be just as

31. Ian Lancashire and Elisa Tersigni, "Shakespeare's Hard Words, and Our Hard Senses," in *Old Words, New Tools: Historicizing Shakespeare's Language in Digital Media*, ed. Jennifer Roberts-Smith and Janelle Jenstad (Ashgate, forthcoming).

32. Ian Lancashire, "William Cecil and the Rectification of English," in *The Languages of Nation: Attitudes and Norms*, ed. Carol Percy and Mary Catherine Davidson (Bristol: Multilingual Matters, 2012), 39–62.

33. Claudius Hollyband, *A Dictionarie French and English* (London: T. O. for Thomas Woodcock, 1593), sig. E6r.

34. William Camden, *Remains Concerning Britain*, ed. R. D. Dunn (Toronto: University of Toronto Press, 1984), 54, 80.

35. A neologism of my own, occasioned by respect for the intellectual achievements of early lexicographers: Ian Lancashire, gen. ed., "Series Preface," *Ashgate Critical Essays on Early English Lexicographers*, 5 vols. (Farnham, Surrey: Ashgate, 2012), xi. For analyses of early dictionaries as lexicature (lexical

fascinating to curious readers today as the imaginative literature of the period's great writers.

Three factors account for the minimal overlap between the *OED Online* and *LEME*. Oxford lexicographers have said that they are disinclined to include, among illustrative quotations, word-entries by pre-modern lexicographers, that is, those who did not build chronological, regional, and subject dictionaries to the scientific standards employed by the *OED Online*. Its website explains that it "look[s] for examples of uses of a word that are not immediately followed by an explanation of its meaning for the benefit of the reader."[36] This view holds that lexicographers who write a dictionary definition for a word are not themselves using that word contextually.[37] The second justification for the limited use of early dictionary word-entries as quotations by the *OED Online* is that old lexicographers were thought to plagiarize from one another. Both points appear in an article by James Riddell in 1974.[38] Drawing on Starnes and Noyes's source studies of Latin-English dictionaries, Riddell warns that words cited only in Early Modern English dictionaries include no "actual instance of […] use." He adds that they are also unreliable because they invented words, plagiarized, and often erred in assigning meaning. The third factor in reluctance by the *OED Online* to quote pre-modern dictionary definitions is the unavoidable repetition that citations of earlier dictionary explanations would occasion.[39]

Jesse Sheidlower, an *OED Online* lexicographer, explains the objection from the perspective of usage:

> The best quotes are contextual, unexplained examples. These show that the word is really in use, and they show how it is in use. The worst are glossarial quotes, which can, if the source is trustworthy, be helpful to the

---

literature), see Gabriele Stein, *John Palsgrave as Renaissance Linguist: A Pioneer in Vernacular Language Description* (Oxford: Clarendon Press, 1997), and *Sir Thomas Elyot as Lexicographer* (Oxford: Clarendon Press, 2014).

36. See *OED Online*, 20 November 2014, http://public.OED.com/about/frequently-asked-questions/ and under "How does a word qualify for inclusion in the *OED*?"

37. The *OED Online*, however, will always cite an early dictionary quotation when it is the earliest one available.

38. James A. Riddell, "The Reliability of Early English Dictionaries," *The Yearbook of English Studies* 4 (1974): 1–4.

39. Peter Gilliver, oral communication, 24 May 2013.

> lexicographer in writing the definition, but which are not generally very useful to quote. Similar examples are also to be avoided, such as a use that is in running text, but which is followed by a definitional or explanatory phrase, or a use which is in scare quotes, indicating that the term is thought to be new, unnatural, non-serious, etc. (203)[40]

This objection to glosses or definitions as quotations is possibly a red herring because most word-entries by early modern English glossographers do not annotate words by defining their meaning. Being overwhelmingly bilingual or polyglot, they give English equivalents or synonyms for foreign-language terms. The glossographer selects the English equivalent specifically because it requires *no* explanation: the English word or phrase belongs to the simple mother tongue that few native speakers would fail to understand. Why would a bilingual lexicographer require a reader to fetch a separate monolingual hard-word dictionary before seeing how to translate a foreign-language word? However, Sheidlower's position makes sense for a dictionary like the *OED Online*.

A gloss resembles a pairing of two synonyms in the form *a is b.* Sheidlower concedes some value for glossarial quotations by saying that 3 to 4 percent of the total quotations of the *Middle English Dictionary*, a production that the *OED Online* rightly considers good, are glossarial. Why then does the *OED Online* not favour glosses and word-explanations in early dictionaries in quotations? Sheidlower helpfully says that "the *OED* confines itself to describing the semantic development of a word from quotation contexts; the *MED* considers facts about language use besides contextual semantics, such as sources and text types, as part of lexical history."[41] Thus the *OED Online* serves lexical history but does not commit itself to include all the source materials, the "facts" (such as what a single person believes a word's meaning to be) that researchers might find valuable. As the *OED Online* web site says today, it focuses on "the 'real' facts of the language." *MED* has a relaxed inclusion policy, and *LEME*, which refuses to define any word itself, a plainly extreme one: it aims to include all "contemporary comments" on English words.

40. Jesse Sheidlower, "How Quotation Paragraphs in Historical Dictionaries Work: The *Oxford English Dictionary*," in *Contours of English and English Language Studies*, ed. Michael Adams and Anne Curzan (Ann Arbor: University of Michigan Press, 2011), 191–212.

41. Sheidlower, 204.

If a post-lemmatic field does not explain word-meaning, as I believe, what does it do? Although some definitions do appear in Early Modern English, as in medical and legal lexicons, they normally explain *the thing* a word denotes rather than the lexical connotation of the headword itself. Even Samuel Johnson did not register a lexical sense of the definition in his own dictionary. The late eighteenth-century marked what I call the great definition shift, when the emerging sciences, and philosophy, shifted the object of a definition from a thing to the word denoting it so as to ensure that thinkers knew what they were talking about.[42] Logical definitions—as classical rhetoric interpreted them, of *things*—gave way to lexical ones, and of course we and the *OED Online* are of this second era. Up to the end of the early modern period, speakers thought of language in a fundamentally different way than we have in the past two hundred years. For text before 1700, then, should a period dictionary not explain words where possible with reference to the things they denote, or to the foreign-language terms they were said to translate? If we are to understand language used in early modern works, we should have the option of seeing it in terms of the language theory of the early moderns. *LEME* gives us that option and does not exist to compete with or question the research agenda of Oxford lexicographers.

Scholarship on the charge of plagiarism levelled against the early glossographers has advanced since Riddell. Jürgen Schäfer surveyed hard-word glossaries and dictionaries from 1475 to 1640 and said: "They were not merely indiscriminate copiers but true pioneers in the field of lexicography, worthy ancestors of Samuel Johnson."[43] Recently, Roderick McConchie has redeemed even Milton's nephew Edward Phillips, whom Thomas Blount attacked for plagiarizing his *Glossographia* (1656).[44] McConchie argues that any lexicographer may adopt the wording of another for good reasons. For instance, why change what is efficient and correct? Yet the honesty of English lexicographers is acknowledged to be excellent. Louis Cooper as early as in 1962 established that,

42. Ian Lancashire, "'Dumb Significants' and Early Modern English Definition," *Literacy, Narrative and Culture*, ed. Jens Brockmeier, Min Wang, and David R. Olson (Richmond, Surrey: Curzon, 2002), 131–54.

43. Jürgen Schäfer, *Early Modern English Lexicography*, 2 vols. (Oxford: Clarendon Press, 1989), 1:8.

44. R. W. McConchie, "'The Most Discriminating Plagiarist': The Unkindest Cut (and Paste) of All," in *Selected Proceedings of the 2012 Symposium on New Approaches in English Historical Lexis*, ed. R. W. McConchie, T. Juvonen, M. Kaunisto, M. Nevala, and J. Tyrkkö (Somerville, MA: Cascadilla, 2013), 107–19.

unlike their Spanish counterparts, most English glossographers, John Minsheu included, acknowledged their sources.[45]

The current *OED Online* is sensitive to some of these issues. Which perspective we have "on the 'real' facts of the language" will depend on what we are looking for: a mirror to early modern thought, or the best research on the vocabulary of the English language.

LEME now

LEME supplies a valuable and neglected dataset (as noted in David Vancil's review; see note 1), but unlike almost all comparable tools today is constrained by being the responsibility of a single person for nearly thirty years. After retiring in mid-2012, I have worked on *LEME* five days a week. At first I populated *EMEDD* with texts by relying on individual SSHRC grants, but for the past eight years it has been funded entirely by the royalties received from *LEME* licensing and by the university's work-study program.[46] Despite this, *LEME* has grown by 20 percent, thanks to data-input by a small number of loyal student assistants. I have also benefitted from the wonderful resources of Early English Books Online/Text Creation Partnership (EEBO/TCP), which has transcribed some large dictionary texts of the period.

My California colleague twenty-five years ago and reviewers of my grant applications justifiably warned me about the audaciousness of this task. I have just finished proofing John Rider's *Bibliotheca Scholastica* (1589), which has over forty thousand entries and subentries, English and Latin. It took four people—including a typist in Beijing who entered from the middle of letter C to the end—over five years to draft the text-entry. My proofing went on intermittently for two more years. I have had regular trouble distinguishing between its *n* and *u*, *f* and long-*s*, *c* and *e*, and *t* and *r*. Some 1,700 times I had to emend for

45. Louis Cooper, "Plagiarism in Spanish Dictionaries of the XVIth and XVIIth Centuries," *Hispania* 45.4 (Dec. 1962): 717–20.

46. I will mention two exceptionally hard-working *LEME* assistants, Sarah Greene and Ruth Peidi Zhao. Greene undertook the difficult transcription of several dozen dictionaries and glossaries over five years. Zhao, a specialist in French linguistics, entered Claudius Hollyband's French-English dictionary of 1593, and the 25,500 word-entries of a hard Latin-English dictionary, *Ortus Vocabulorum* (1500), which is obscured by a dense layer of abbreviations, and is co-editing with me the *Universal Etymological English Dictionary* published by Nathan Bailey in 1737.

the printer's failure to insert a soft hyphen. Encoding the text also proved difficult, particularly the need to specify, for every word, its language. That trained Latinists may find some things to complain about lies at the editor's feet, but they should remember John Rider's wisdom when he cut to the quick of the average retired professor: "An olde dog past the best. *Canis emeritus*."

Although anyone can feel overmatched by dictionary data at times, they are not big data. Even *OED Online* word-entries are actually small and well-formed enough to be used as queries on big data: standard headwords spelled in modern English, linked to customary old spellings, etymologies, senses, and quotations, can be fitted into a query so as to locate a manageable body of hits in the Google-sphere. The thesaurus now attached to the *OED Online* also makes possible a query not only for one headword but for its synonyms and antonyms as well. Had the Poetry Foundation of Chicago not paid *Representative Poetry Online* (*RPO*) at Toronto[47] a substantial fee to license its database in perpetuity, I would not have seen the potential value of smaller humanities databases to big data. It was partly the SQL structure that made *RPO* so useful to Chicago. The University of Toronto found that a free poetry project was also an OEM.[48] *LEME*'s usefulness too may be more than lexicographic, if its lemmatized, modernized headwords help interrogate bigger data-sets like the forty thousand searchable texts in the EEBO/TCP database or the more than thirty million scanned books in the Google Books collection. If these were lemmatized, consider how different the results of searches on them would be. Enabling such lemmatization has become an operational goal for *LEME*. In a related study of the vocabulary size of letter-writers and diarists of the period, Elisa Tersigni and I are developing a lemmatizing system for early modern texts that, at present, has a dictionary of twenty-eight thousand spellings with modern lemma and part of speech.[49]

Canadian web databases that serve the early modern English period, such as the Internet Shakespeare Editions (ISE; 1996–), *Iter* (1994–), *Orlando* (2006–),[50] and *LEME*, emerged at a juncture in scholarship. Until the 1990s,

47. Ian Lancashire, Marc Plamondon, and others, eds., *Representative Poetry Online* (Toronto: University of Toronto Libraries, 1994–), http://rpo.library.utoronto.ca.

48. Original Equipment Manufacturer.

49. Ian Lancashire and Elisa Tersigni, "Early Modern English Vocabulary Growth," in *Corpus Linguistics 2013: Abstract Book* (Lancaster: Lancaster University, 2013), 156–59.

50. Susan Brown, Patricia Clements, and Isobel Grundy, eds., *Orlando: Women's Writing in the British Isles from the Beginnings to the Present* (Cambridge: Cambridge University Press, 2006–), http://orlando.

researchers undertook article- or book-sized products to be published by a press. Then computers arrived, encouraging really ambitious projects. Formerly, any work with a scope larger than an individual could manage, such as a complex edition, a series, or a journal, was a collaboration orchestrated by a general editor, a project leader, or a publisher. The first generations of web databases seemed to fit into this model: Broadview Press collaborated with ISE, Cambridge University Press with *Orlando*, and the University of Toronto Press with *LEME*. The example I offer, however, might well dissuade researchers from undertaking data-intensive projects that deep pockets or crowd-sourcing are better equipped to produce: library consortia (Internet Archive), big business (Google Books), and the gifted collective that quickly amassed the greatest information resource on the planet (Wikipedia). Doubtless, a business would have input *LEME* faster and more efficiently than I have, but it would not have done so because the market was too doubtful. And there is always some merit in putting a researcher in charge who would observe, for example, how different the language theory and practice of Early Modern English was from our own. Without thirty years of encoding and proofing, I might not have noticed the lack of lexical definition, the English obsession with hard words, native indifference to the English tongue, and the power of patrons to effect language change.

Ultimately researchers hope to add new work to the lasting architecture of knowledge. In one perspective, big data serves researchers magnificently; in another, it colonizes them. Big data popularizes research well suited to big data.[51] The more text databases that we create and must maintain to contribute to big-data scholarship, the greater the budgetary strain both makers and libraries experience. Responsibility for a book ends at publication, but online researchers expect that online scholarly big-data resources should grow in size and accessibility, long after the original publication date, with the Internet itself: what would once have been a book becomes, by default, a costly "journal" that a single scholar or project must populate indefinitely. The history of large Canadian scholarly web enterprises like *LEME* hints at the value of thinking strategically

---

cambridge.org; *Iter* (Toronto: University of Toronto Libraries, 1997–), http://www.itergateway.org/; *Internet Shakespeare Editions* (Victoria, BC: University of Victoria, 1996–), http://internetshakespeare.uvic.ca/.

51. The Digging into Data Challenge by SSHRC, NSERC, CFI, the Institute of Museum and Library Services, NEH, NSF, and other national agencies is an example: http://www.diggingintodata.org (accessed November 20, 2014).

about how our research will benefit online resources in the future.[52] Should we forge a new collective will to collaborate and compete with commercial big data, as Wikipedia has? Should individuals still be encouraged to undertake projects that require large-scale and ongoing data creation when funding resources are diminishing? The online agora no doubt has surprises in store for us yet.

52. A reviewer asks, "Will *LEME* continue without Lancashire?" The answer to that question will come from the recommendations of its academic advisory board. They will undoubtedly pay close attention to what *LEME* users in the future—licensed and unlicensed—want and to the ability of the web at that time to supply it.