

Documenting the Final Days of Monolingual English Learners' Dictionaries Using the Archived Web

Geraint Paul Rees

Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona

E-mail: geraintpaul.rees@upf.edu

Abstract

Online dictionaries have many advantages over their physical counterparts. However, the ephemeral nature of web content means that they are often changed without notice and no ostensible record of what came before remains. This makes research on historical online dictionaries difficult and perhaps explains why, while the history of printed monolingual English learners' dictionaries (MELDs) has been comprehensively explored, studies of online dictionaries have tended to take a cross-sectional rather than longitudinal view. This is not ideal since it means that a large period of MELD history is yet to be explored. Moreover, given recent predictions of the decline of MELDs, as we know them, in light of developments with AI chatbots and other digital tools, this gap is all the more significant. In an attempt to remedy this situation, this study applies Brügger's (2018) framework for archived web research to explore the feasibility of using the web archive, *the Wayback Machine*, to trace the development of websites that give, or have given, access to 'the big five' MELDs. Some key challenges of using archived web material to conduct lexicographic research are discussed along with suggestions for potential solutions.

Keywords: digital dictionaries; monolingual English learner's dictionaries (MELDs); web archives; Internet Archive; history of lexicography

1. Introduction

The transition from paper-based to online dictionaries brought an abundance of new opportunities for lexicographers and lexicography researchers. For example, methods such as log-file analysis and eye tracking have allowed greater insight into the look-up behavior of users, while techniques from Human Computer Interaction research have been applied in an attempt to improve usability (Frankenberg-Garcia et al., 2019). Moreover, the increased variety of formatting options afforded by online dictionaries has led to greater consideration of visual concerns such as digital typography (Dziemianko, 2015; Hao et al., 2022), and, more generally, of the efficient visualization of information in ubiquitous lexicographic resources (Rees, 2022) such as writing assistants (Roberts et al., 2020). It has also brought new phenomena for investigation, for example, advertisements in online dictionaries (Dziemianko, 2019, 2020). However, this research has typically employed cross-sectional rather than longitudinal designs. The historical evolution of online dictionaries has been largely ignored.

This dearth of historical research may be the result of a scarcity of historical data with which researchers can work. While comparing different editions of printed dictionaries can reveal changes in their makeup over time, digital editions are ephemeral; once changes have been made to a dictionary website, records of the previous version are

often, ostensibly, lost. From the users' perspective, this transitoriness has been reported as undermining confidence in digital dictionaries (Dziemianko, 2018). From a historiographic perspective, it is also problematic. Monolingual English learner's dictionaries (henceforth MELDs) are a good case in point. Most historical overviews have focused on printed versions of these resources. This is understandable in the case of Cowie (1999) who writing in the early days of the web provided a thorough historical treatment of paper-era MELDs—from the pioneering resources of the 1930s and 40s, to the second generation of the 1960s and 70s, and the third generation created using machine-readable corpora in the 1980s and 1990s. However, more recent historical accounts of English learner's dictionaries have also focused on printed editions. For example, Yamada and Xu (2024:109-10) focus on printed dictionaries in their historical overview of English dictionaries for learners with the justification: “nearly the entire history of learner's dictionaries is that of print dictionaries”. While the spirit of the statement—that there is a great deal of printed MELD material to investigate and a great deal to be gained from doing so—is undoubtedly true, from the point of view of historiographic thoroughness the exclusion of online resources is questionable.

Taking a liberal interpretation, MELD origins could be dated 100 years back in time to the Vocabulary Control Movement of the 1920s and 30s, while the emergence of online MELDs could be dated to around the turn of the last millennium. This leaves a gap (almost a quarter of a century) that comprises around a quarter of total MELD history in need of analysis. Even with more conservative assumptions about MELD origins offline and online, the gap in historical treatment remains significant. If, as is often claimed, the publication of *The Idiomatic and Syntactic English Dictionary* in 1942 (some 83 years ago) marks the birth of the MELD as a genre (Cowie, 1999) and its descendant the *Oxford Advanced Learners' Dictionary* was first made available online in 2010 (some 15 years ago), the OALD has had an online presence for almost one fifth of its history.

Whatever the precise extent of the period in which MELDs are missing historical analysis, it clearly represents the most recent period in their history. Crucially, in light of recent developments with AI chatbots and other digital language tools (De Schryver, 2023; Lew, 2024; Nesi, 2024; Rees & Frankenberg-Garcia, 2025a) and the shuttering of the highly acclaimed online *Macmillan English Dictionary* at the end of June 2023, these may well be the final days of MELDs as we know them. In this context, the importance of the preservation of historical data from online MELDs is clear. In lexicography more generally, Preston-Kendall (2025) discusses the causes and consequences of the lack of preservation of previous versions of online dictionaries. The causes include market forces and political pressures. The consequences include undocumented changes that have negative implications for scholarship, as well as, in some cases, the disappearance of resources from the web completely. The loss of technological knowledge (i.e. knowledge about how to produce dictionaries) could also be added to this list (Rees, 2025a).

To accomplish the goal of preserving historical data from MELDs, we first need to know what records of online MELDs in their past states exist. With this in mind, this exploratory paper aims to establish the extent of the MELD material available in a popular web archive, *The Wayback Machine*. This material will be used to provide a timeline of key dates in the evolution of MELDs online and to reflect on some of the possibilities and challenges of using archived web material to conduct lexicographic research.

1.1 Digitality, Web Strata, and Online Dictionaries

To carry out historical research using archived web material, it is necessary to understand the unique characteristics of this material as an object of research. Furthermore, in the context of this study, it is necessary to understand how these characteristics relate to online dictionaries. In literary theory and critical studies, the notion of ‘textuality’ is used to examine the production, organization, and interpretation of text in the broad sense of a system of signs. Brügger (2018) argues that a parallel notion, ‘digitality’, is needed to examine digital texts. In practice, accounting for the digitality of the web means taking into account that it has both visible (e.g. the interface that the end user sees in the browser) and invisible layers (e.g. the source code of a website); that it is fragmentary (e.g. that webpages are collections of graphics, written text, and videos), and that it is hyperlinked (e.g. that webpages or fragments of webpages can be linked to each other). In addition to the notion of digitality, to help focus web studies, Brügger (2018:30) posits an analytical grid of web strata, “a set of theoretical, systematic subdivisions in manageable and coherent units”. These are shown along with illustrative examples in Table 1.

Stratum	Example at the visible layer	Example at the invisible layer
web element	a video showing the pronunciation of <i>web</i>	the code pertaining to that video e.g. “”
web page	the entry page for the word <i>web</i> in <i>Collins Online Dictionary</i> <collinsdictionary.com/dictionary/english/web>	the code or databases underlying the entry page
website	<i>Collins Online Dictionary</i> <collinsdictionary.com>	the code or databases underlying the site
web sphere	websites providing access to ‘the big five’ MELDs	the code or databases underlying ‘the big five’ websites

Table 1: Brügger’s (2018) web strata

The websites that give access to ‘the big five’ MELDs represent a clear web sphere.

The exploratory analysis in this study takes a necessarily broad perspective focusing on the webpage stratum, although web elements typically used to realize specific dictionary features or encode certain aspects of lexicographic information are also discussed. It would be infeasible for any study, let alone an exploratory one, to analyze all the webpages of a dictionary website. For this reason, the analysis will focus on documenting the presence of archived front pages and entry pages, which represent typical use pathways of dictionary websites (Rees, 2023). It will also record the presence of documentation such as help pages and FAQ pages, potentially allowing researchers to address dictionary use training—a topic of perennial interest in pedagogical lexicography.

2. Method

This section gives more detail about ‘the big five’ MELD websites and how they are understood in this study. It also provides a brief overview of *The Wayback Machine* (WBM) digital archive and its interface. Finally, it outlines a procedure that uses the WBM to explore archived MELD material.

2.1 The ‘Big Five’ MELD Websites

In English learner lexicography the term ‘the big five’ has long been used to refer to the five widely used and researched MELDs produced by major publishers. They are *Cambridge Advanced Learner's Dictionary* (CALD), *Collins COBUILD Advanced Learner's English Dictionary* (COBUILD), *Longman Dictionary of Contemporary English* (LDOCE), *Macmillan English Dictionary for Advanced Learners* (MEDAL), and *Oxford Advanced Learner's Dictionary* (OALD). In short, they are notable for their user-friendliness and features supporting language production. However, the migration from paper-based dictionaries to online dictionaries has blurred the lines between these MELDs and other dictionaries produced by their publishers. Of the ‘big five’ publishers, only Longman and Macmillan offer(ed) direct access to their MELDs (through the URLs [<https://www.ldoceonline.com/>](https://www.ldoceonline.com/) and [<https://www.macmillandictionary.com/>](https://www.macmillandictionary.com/) respectively). Access to the Cambridge (CAM; [<http://dictionary.cambridge.org/>](http://dictionary.cambridge.org/)), Collins (COL; [<https://www.collinsdictionary.com/>](https://www.collinsdictionary.com/)), and Oxford (OX; [<https://www.oxfordlearnersdictionaries.com/>](https://www.oxfordlearnersdictionaries.com/)) products is provided via portals that aggregate content from several different dictionaries by the same publisher. For the average user, the resulting content mix may be beneficial, for lexicography researchers it can create an impression of a patchwork of lexicographic data loosely stitched together. By noting the date at which MELD data was incorporated into the dictionary portal, it is hoped that the timeline produced in this study will help future researchers unstitch the MELDs from the dictionary portal if they so wish. Note that to represent the distinction between portal and dictionary, in the discussion that follows dictionary portals are identified with a different acronym to the specific MELD they contain.

2.2 The Wayback Machine

The Wayback Machine (WBM) (Internet Archive, 2025) is regarded as the most comprehensive web archive publicly available. From its launch in 2001 to July 2025, it had preserved over 946 billion web pages globally, the earliest dating back to 1996. Crucially, given the aims of this study, this includes content that has since been altered or even deleted from the live web. Captures—reproductions of past webpages—are central to the functioning of WBM. They have been gathered by crawlers—programs that navigate the internet copying web content. To create an approximate reconstruction of a website around a given point in time, WBM links captures together with others taken around that point in time. The Calendar view (Figure 1) gives a visual representation of available captures for a given URL on a timeline. By clicking on the circle around a date, users can see captures from that day.

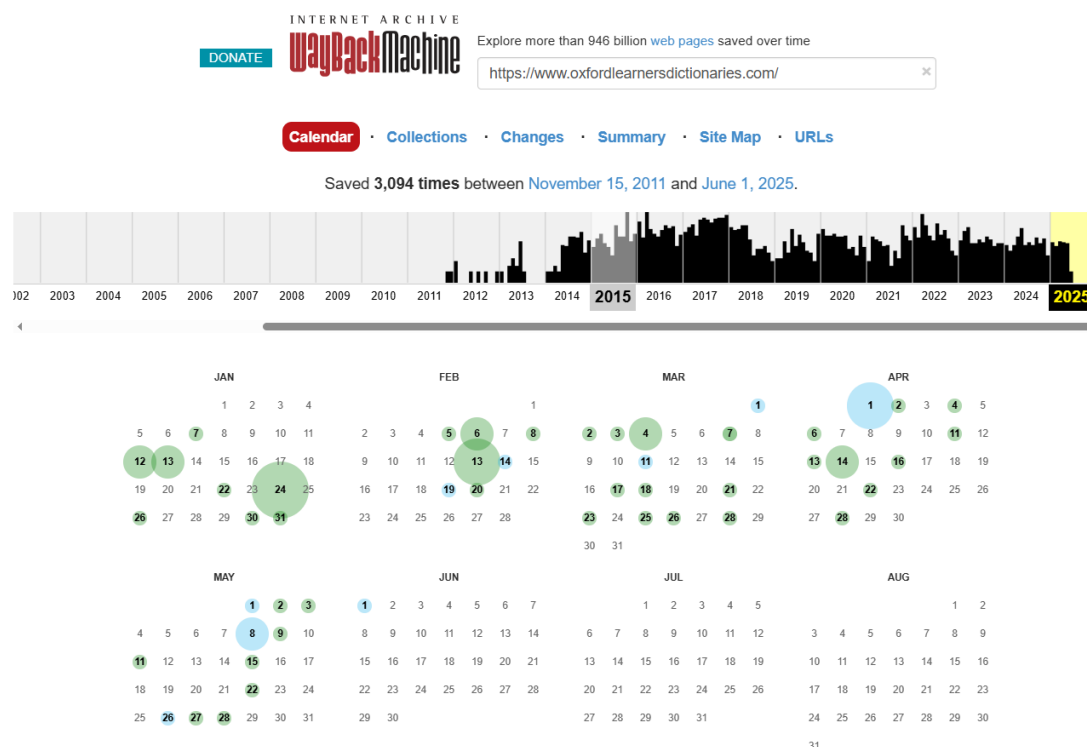


Figure 1: WBM's Calendar view.

2.3 Procedure

The analysis procedure comprises two steps. The first step involves recording the archived web material that is available via WBM for the websites that give or have given access to the 'big five' MELDs. This includes not only establishing the period this material covers but also establishing which elements of the dictionary websites are included in the archived material. As far as temporal coverage is concerned, this is achieved by entering the last known good URL into WBM and navigating to the first capture available. The first capture may provide a representation of the website around

the time it was established. Alternatively, it may contain clues about previous versions of the website under different URLs (e.g. through redirect notices, references in the visible content of the page, or references in its source code). If such clues do exist, the process will be repeated for any relevant URLs found. As far as establishing which elements of dictionary websites are included in the archived material, the front page captures serve as the point of departure. From there, the existence of key elements at the stratum of webpage is recorded (i.e. entry pages and documentation).

The next step involves organizing the available archived material into versions. Versions may be directly acknowledged on dictionary websites, in publishers’ blogs, or marketing material. However, there may be other clues. Following common practice in research on the historical web (Brügger, 2018), similarities and differences in visual appearance, web-design features, and technologies are also used to differentiate versions.

3. Results and Discussion

Table 2 indicates the number of captures available for the front pages of the last known good URL of each resource as of 21 July 2025 and the date of the first and final captures (DD/MM/YYYY). As such, it gives an approximate indication of the extent of archived material available on the online MELDs.

Website	First capture	Final Capture	Total captures
CAM	15/08/2000	17/7/2025	6,722
COL	24/11/2005	8/07/2025	5,258
LDOCE	02/11/2004	21/07/2025	2,457
MEDAL	07/02/2002	01/07/2025	8,758
OX	15/11/2011	08/07/2025	3,104

Table 2: Captures for the front pages of the last known good URL

Taken together, there are 26,299 captures of MELD dictionary front pages spanning from August 2000 to July 2025, a period of almost 25 years. However, these headline figures are somewhat misleading since in many cases there is a significant difference between the date the first capture was taken and the date the MELD was made available online. For example, in the case of CAM, which boasts the earliest online presence of the MELDs examined, the first capture (15/08/2000) postdates the launch of the *Cambridge International Dictionary of English* (a forerunner to CALD) online according to the publisher: “There have now been over **five million** searches on the online dictionaries since they were launched in July 1999” (Cambridge Dictionaries, 2 December 2000 [original emphasis maintained]). Similarly, in the source code in the first capture of CAM, there is a comment: “<!-- end of new bits (10 may 2000) -->” [SIC] suggesting that the site had been online before that date. However, in the majority

of resources examined, the date of the earliest capture tends to overestimate the time MELDs have been available online. For instance, the first capture of COL (24/11/2025) is, in fact, a domain squatter—an early recognition of the commercial potential of online dictionaries. The first two versions of COL proper at <http://www.collinslanguage.com/> evidence access to the general-purpose Collins English Dictionary. It was not until 18 December 2007 that captures indicate the website giving access to the MELD, COBUILD. Similarly, the first captures of MEDAL, are of a “resource site” providing additional material for the paper or CD-ROM resource. Online access to the dictionary proper came in August 2003 and then only for those users who owned the printed or CD-ROM versions. The first capture of OX at <http://www.oxfordadvancedlearnersdictionary.com/> on 18 February 2010 redirects to the publisher’s catalogue entry for OALD7. The first capture of the dictionary proper is dated 1 October 2010. LDOCE is the only resource that has a first capture indicating dictionary access.

Table 3 gives a timeline of online MELD development providing an overview of the major versions noted in the analysis of the website captures. There are clear differences in stability between websites. LDOCE is, and MEDAL was, relatively stable with around 0.2 versions per year. CAM and OLD fill the middle ground with 0.3 versions per year. With 0.4 revisions per year COL, the youngest MELD website, has undergone changes most frequently.

	CAM	COL	LDOCE	MEDAL	OX
Jul-99	Reported launch date				
Aug-00	Capture of V1				
Oct-01	Capture of V2				
Feb-02				Capture of V1 (resource site)	
Sep-02	Capture of V3				
Jun-03				Capture of V2 (resource site)	
Aug-03				Capture of V3 (paywalled)	
Oct-03	Capture of V4				
Nov-04			Capture of V1		
Jul-07		Capture of V1 (collinslanguage.com - paywalled)			
Nov-07		Capture of V2 (collinslanguage.com - CED and CET access)			
Dec-07		Capture of V3 (collinslanguage.com - COBUILD access)			
Sep-08			Capture of V2		
Nov-08			Capture of V3		
Mar-09				Capture of V4	
Feb-10					Redirect to catalogue entry for OALD7

Apr-10					Earliest entry in 'word of the day' RSS feed
Oct-10					Capture of V1 (oxfordadvancedlearnersdictionary.com)
Dec-10	Capture of V5				
Nov-11		Capture of V4 (collinsdictionary.com - marked as 'beta')			
Jun-12		Capture of V5 (no longer beta)			
May-14					Capture of V2 (oxfordlearnersdictionaries.com)
Feb-15					Capture of V3
Jun-16		Capture of V6			
Jul-16	Capture of V6				
Sep-16			Capture of V4 (similar to published version of 21/07/2025)		
Sep-19	Capture of V7 (similar to published version of 21/07/2025)				
Nov-19		Capture of V7 (entry and index pages blocked)			
Jan-20				Capture of V5 (similar to final capture of	Capture of V4 (similar to published version of 21/07/2025)

				dictionary website dated 01/07/2023)	
May-24		From here on crawlers are blocked. The published version of 21/07/2025 is different from the last good capture.			

Table 3: Key moments in the development of the ‘big five’ MELD websites

3.1 Key Challenges

The creation of the timeline outlined in Table 3 was not without challenges. This section briefly discusses a few of the issues encountered in the capture analysis before going on to discuss some potential solutions. Further details about the captures collected for analysis including hyperlinks to the captures themselves can be found in the online supplementary material <<http://bit.ly/4532IeF>>.

3.1.1 Fuzzy Boundaries between Versions

At times, website versions are explicitly acknowledged, either at the visible layer—for example, “A Message from the Editor” of CAM announcing changes (11/04/2010) or COL (Version 4) acknowledging a beta version—or at the invisible layer in a source code comment (e.g. CAM, Version 1 discussed above). However, more often than not, changes are not explicitly acknowledged, there is simply a notable change between one capture and the next. As noted in Brügger (2018:145), visual appearance, web-design trends, and web technologies can be used to differentiate versions. However, the question of where to draw the line is often not clear.

During the analysis of captures, it was striking that features that are present in one version often disappear in the next only to reappear in future versions. For example, help pages disappear in Version 4 of COL but reappear in Version 9. In this study, this along with other primarily visual changes was judged to be sufficient evidence for a major version change. Minor versions, are those where features disappear only to reappear in captures that are broadly similar in all other ways a few months later. This occurs with captures representing CAM where the site’s alphabetical index disappears in Version 5.1 (31/01/2011) yet appears again in Version 5.3 (26/05/2011). This is testament to the ephemeral nature of web content and represents a challenge when conducting historical research using archived online dictionaries. Ultimately, in future historical studies of lexicographic web resources, the degree of granularity in recording should depend on the precise research question.

3.1.2 Dialect Specific Versions

The existence of different versions of resources for different varieties of English (in this case British English vs. American English) also poses problems when trying to establish the extent of the archived material available on MELDs. Many of the captures reported in the WBM calendar view do not actually contain reconstructions of the website at the target URL but instead redirect to the nearest capture of a localized version. For example, from 2015 onwards captures of <<http://www.oxfordlearnersdictionaries.com/>> redirect to the US version

<<http://www.oxfordlearnersdictionaries.com/us/>>. This issue generally affects later versions of the resources. The target of the redirection depends on the geographical location of the server that made the initial crawl. Most of the collections of crawled websites employed by WBM were made by organizations in the United States. Whether the difference in provenance in the archived material limits its usefulness for investigating historical web dictionaries ultimately depends on the scope of the research question.

3.1.3 Inconsistent and Incomplete Coverage

Figure 2 compares the WBM coverage of front pages for CAM (upper) and LDOCE (lower), the two resources that have had the same URL over time. It not only illustrates periods that are not covered in the archives, but also the striking difference in the temporal coverage of the archived material across dictionary websites. There are long periods of time for which no captures of LDOCE are present in the archived material. Coverage of CAM, in comparison, is more comprehensive.

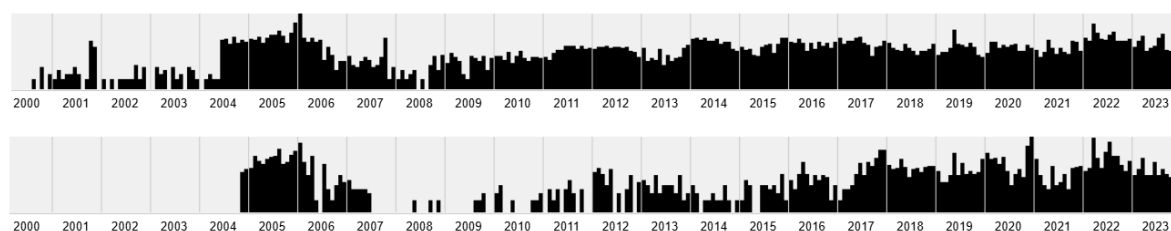


Figure 2: WBM coverage of front pages for CAM (upper) and LDOCE (lower)

Gaps in coverage are even more pronounced beyond the dictionary front pages. For example, although all archived versions of the MELDs examined contain entry pages for some words, coverage is patchy. Due to the way in which crawlers navigate the web, it is typically entry pages for those words that have appeared on the front page of the dictionary in ‘word of the day’ or ‘new word’ features that are archived. The entry pages for most words have not been archived. Moreover, a given word may have been archived for one version of the website but not for another. Similarly, since search functionality is not available on most archived websites one cannot simply navigate to the entry page for a word of interest using the search box. Obviously, this could impede the comparison of entries for a given word between historical versions of different dictionaries. It could also impede the comparison of the entries for a given word for different versions of the same dictionary.

However, it may be possible to discern general differences in elements of entries or differences in the entry for a given word between some, but rarely all, versions. For example, Figure 3 shows part of the entry for *table* (noun)—typical of polysemous noun entries—from version 2 of OLD (29/03/2014). Figure 4 shows part of the entry for the version of OLD that was live on 03/10/2025. Red, uppercase lettering is used for the

semantic category label ‘furniture’ in the older version, while lowercase, grey letters are used in the newer one. The use of color and uppercase lettering may have been carried over from the paper version of the dictionary, where tighter space limitations meant that greater typographical variation was needed to distinguish between entry parts.



table *noun*
 'teɪbl̩ ; 'teɪbl̩

FURNITURE

1 a piece of furniture that consists of a flat top supported by legs

- ◆ a *kitchen table*
- ◆ A *table for two, please* (= in a restaurant).
- ◆ I'd like to **book a table** for tonight (= in a restaurant).
- ◆ to **set the table** (= to put the plates, knives, etc. on it for a meal)
- ◆ (British English also) to **lay the table**
- ◆ to **clear the table** (= take away the dirty plates, etc. at the end of a meal)
- ◆ He questioned her next morning over the *breakfast table* (= during breakfast).
- ◆ (British English, formal) Children must learn to behave **at table** .
- ◆ a *billiard/snooker/pool table*

HELP There are many compounds ending in **table**. You will find them at their place in the alphabet.

Search Results

table noun

- table verb
- multiplication table noun
- end table
- top table noun
- table mat noun
- bird table noun
- head table noun
- high table noun
- card table noun
- round-table adjective
- times table
- table lamp noun
- table linen noun
- table wine noun
- water table noun
- bedside table noun
- coffee table noun
- dining table noun
- dinner table noun

table *noun*

A1

/'teɪbl̩/

/'teɪbl̩/

Idioms

furniture

1 **A1** a piece of furniture that consists of a flat top supported by legs

- at a/the table *We sat at a round table in the corner.*
- around/round a/the table *They were sitting around the kitchen table.*
- She took a seat at the end of the table.
- My father always sits at the head of the table.
- A table for two, please (= in a restaurant).
- I'd like to **book a table** for dinner tonight (= in a restaurant)
- to **set the table**
- (British English also) to **lay the table** (= to put the plates, knives, etc. on it for a meal)
- to **clear the table** (= take away the dirty plates, etc. at the end of a meal)
- He questioned her next morning over the *breakfast table* (= during breakfast).
- a pool/billiard/snooker table

SEE ALSO **bird table**, **card table**, **changing table**, **coffee table**, **dining table**, **dinner table**, **head table**, **high table**, **top table**



Other results

All matches

table *verb*

table

end table *noun*

table mat *noun*

top table *noun*

bird table *noun*

card table *noun*

head table *noun*

See more

+ Idioms

Nearby words

tabla *noun*

tablature *noun*

table *noun*

table *verb*

tableau *noun*

3 October 2025



chilly

Figure 3: Part of the entry for *table* (noun) from Version 2 of OLD (29/03/2014)

Figure 4: Part of the entry for *table* (noun) from the version of OLD live on 03/10/2025

In addition to the increased number of usage examples in the newer version, there are other innovations. For example, a prominent ‘idioms’ button that takes users directly to the relevant section of the entry. This is perhaps representative of a general trend towards greater focus on phraseology in learner lexicography over recent decades (Rees, 2021). Similarly, the inclusion of a label indicating the CEFR level of the headword (in this case A1) aligns with an increasing interest in the lexicographic applications of the CEFR (Lew & Wolfer, 2024; Wolfer & Lew, 2025). The addition of a photograph showing a table and chairs in a dining room illustrates a trend towards increased use of photographs in dictionaries, even though empirical studies indicate that schematic images such as line drawings are often more helpful (Rees, 2025b). Finally, while the older version refers users looking for compound forms to the dictionary’s alphabetical index, the newer version provides direct hyperlinks to these forms. De Schryver (2003) reported that liberation from the constraints of the alphabetical index was already a widely acknowledged advantage of electronic dictionaries. However, as Figure 3 shows, it is one that was still waiting to be put into practice in a major MELD over 10 years later. Indeed, the alphabetical index remains the only way to find some derived wordforms in several MELDs (Rees, 2024).

Nonetheless, even the possibility of examining general differences between dictionaries and versions could be hindered by the fact that occasionally webpage captures are incomplete. For example, the CSS file responsible for formatting the text is missing for some entry pages.

More generally, even sporadically archived captures can show the practical beginnings of what are currently considered novel trends in lexicography research. For instance, early versions of MELD front pages (e.g. COL, Version 1, 06/07/2007) contain references to browser plug-ins designed to provide in situ definitions on webpages. These are redolent of the increasing interest in research on writing assistants and e-readers in recent years (Fuertes-Olivera & Tarp, 2020; Rees & Frankenberg-Garcia, 2025b). Similarly, publisher awareness of web accessibility (making websites accessible to anyone irrespective of disability or impairment), an emerging trend in lexicography research (Arias-Badia & Torner, 2023; Rees, 2023, 2025c), can be seen in OX Version 3 onwards (01/02/2015).

There is some evidence that publishers’ attempts to protect their intellectual property so that it cannot be easily scraped or used for AI training has led to less complete archive coverage. These attempts have had the unfortunate effect of stopping the dictionary material from being accessed through crawlers used by WBM. This is apparent in the case of COL from Versions 7 and 8 where crawling of entry and index pages has been blocked and from 11 August 2024 onwards when crawlers’ site access is blocked completely. Detailed discussion of the legality of web archiving is beyond the scope of this paper. It depends not only on the jurisdiction of the copyright holder but

also on that of the archiver. However, the ability for content owners to opt out and have material removed from archives under certain circumstances has been used to justify web archiving for historical preservation as fair use (Hirtle, 2003).

A key part of the MELD commercial ecosystem and an established object of online dictionary research (Dziemianko, 2019; 2020), advertising is conspicuous by its absence from the visible layer of the archived dictionary websites examined. However, a look at the source code reveals references to online advertising platforms suggesting that the archived websites did contain advertisements when live.

3.2 Solutions

Although it is far from the ideal source of historical data, there is a great deal of lexicographic research that can be conducted using existing archived material on online MELDs. Even with the gaps in temporal coverage discussed, it has been possible to sketch a picture of the evolution of the resources through time.

Future studies could employ convenience sampling of the available archived entry pages. This would allow the investigation of almost all elements of microstructure. These include typography, pronunciation (this would be primarily transcription, but OX captures did contain archived audio files), usage information, entry navigation devices, and even pictorial illustrations and video. However, the analysis of hyperlinks between entries, a key advantage of online dictionaries and a key component of the digitality of the web in general, would be limited by the scope of the sample.

In historical studies of the web in general, screenshots have been used where no fuller-featured archived web content is available. For example, in his history of Facebook, Brügger (2015) relied on screenshots of the social network since no WBM material was available. In lexicography, there are many research papers containing historical screenshots that could give researchers insight into how resources were in the past. Video screen captures could also be used to fill gaps in archived materials. These have the added advantage of being able to capture user interaction with websites. As such they could be used to mitigate the missing interactive features in archived dictionary websites (e.g. search and spelling suggestions). Beyond lexicography, screen captures have been used in historical studies of advertising on the web (Jessen, 2010). This suggests that they could be used to record advertising on dictionary websites; an element that was missing from the captures of MELDs.

Expecting MELD publishers, many of whom are suffering commercial pressures, to make wholesale changes to their resources to facilitate their historical preservation is probably not a realistic hope. They would have little to gain from preserving previous versions of entries online. On the contrary, the bad publicity and loss of advertising revenue that could result from an old entry with politically incorrect content is a real danger. Although, as one reviewer of this paper suggested, consistent entry timestamps

at the visible layer, in addition to those that already occasionally exist in an entry page's source code, might help future researchers date archived entries. Moreover, this would have the added benefit of giving end users an indication of the recentness of the entry they are viewing.

4. Concluding Remarks

This exploratory study has taken a broad-brush approach to establishing the extent of the archived versions of the 'big five' MELDs available in the Wayback Machine archive. The decision to consider this whole web sphere rather than focusing on a specific resource has prioritized breadth of analysis over depth. Nonetheless, establishing a comparative timeline is an important preparatory step for future historical research on archived MELDs that could examine the historical evolution of specific aspects of these resources in more detail. For a finer-grained analysis of the archived material available on historical versions of OLD see Rees (2025a).

The study has considered material from one web archive, the Wayback Machine. This is not only because it is the most comprehensive publicly accessible web archive but also due to the limited time available to conduct analysis of other resources and because since the MELDs analyzed have *.com* domains they are excluded from many national web archives. However, national web archives may be a fruitful source of material for researchers who wish to study the history of online national dictionaries.

Irrespective of the precise nature of the archived material used. It is hoped that this paper will provoke debate on the historical preservation of online dictionaries and promote future studies of their history.

5. Acknowledgements

The research reported here was supported in part by the research project 'Accessible Digital Language Resources: Building an Inclusive Society' (ADLAR) (PID2024-157523OB-I00). I am very grateful for the feedback and encouragement of the two anonymous reviewers of this paper.

6. References

- Arias-Badia, B., & Torner, S. (2023). Bridging the gap between website accessibility and lexicography: Information access in online dictionaries. *Universal Access in the Information Society*, 23, 545–560. <https://doi.org/10.1007/s10209-023-01031-9>
- Brügger, N. (2015). A brief history of Facebook as a media text: The development of an empty structure. *First Monday*, 20(5). <https://firstmonday.org/ojs/index.php/fm/article/view/5423>
- Brügger, N. (2018). *The Archived Web: Doing History in the Digital Age*. The MIT

- Press. <https://doi.org/10.7551/mitpress/10726.001.0001>
- Cambridge Dictionaries. (2000, December 2). *All-time top 20*. <https://web.archive.org/web/20001202072500/http://dictionary.cambridge.org/top20.htm>
- Cowie, A. P. (1999). *English Dictionaries for Foreign Learners: A History*. Clarendon Press. <https://global.oup.com/academic/product/english-dictionaries-for-foreign-learners-9780198235064>
- De Schryver, G. (2003). Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography*, 16(2), 143–199. <https://doi.org/10.1093/ijl/16.2.143>
- De Schryver, G.-M. (2023). The future of the dictionary. In E. Finegan & M. Adams (Eds.), *The Cambridge Handbook of the Dictionary*. Cambridge University Press.
- Dziemianko, A. (2015). Colours in online dictionaries: A case of functional labels. *International Journal of Lexicography*, 28(1), 27–61. <https://doi.org/10.1093/ijl/ecu028>
- Dziemianko, A. (2018). Electronic dictionaries. In P. Fuertes-Olivera (Ed.), *The Routledge Handbook of Lexicography* (pp. 663–683). Routledge.
- Dziemianko, A. (2019). The role of online dictionary advertisements in language reception, production, and retention. *ReCALL*, 31(1), 5–22. <https://doi.org/10.1017/S0958344018000149>
- Dziemianko, A. (2020). Smart advertising and online dictionary usefulness. *International Journal of Lexicography*, 33(4), 377–403. <https://doi.org/10.1093/ijl/ecaa017>
- Frankenberg-Garcia, A., Lew, R., Rees, G. P., Roberts, J., Sharma, N., & Butcher, P. (2019, September). *Collocations in e-Lexicography: lessons from Human computer interaction research* [Workshop presentation]. Pre-conference workshop on collocations at eLex 2019, Sintra.
- Fuertes-Olivera, P. A., & Tarp, S. (2020). A window to the future: Proposal for a lexicography-assisted writing assistant. *Lexicographica*, 36(2020), 257–286. <https://doi.org/10.1515/lex-2020-0014>
- Hao, J., Xu, H., & Hu, H. (2022). A multimodal communicative approach to the analysis of typography in online English learner's dictionaries. *International Journal of Lexicography*, 35(2), 234–260. <https://doi.org/10.1093/ijl/ecab031>
- Hirtle, P. B. (2003, November). *Digital Preservation and Copyright*. Stanford University. https://fairuse.stanford.edu/2003/11/10/digital_preservation_and_copyr/
- Internet Archive. (2025). *Wayback Machine*. <https://archive.org/web/>
- Jessen, I. B. (2010). The aesthetics of web advertising: Methodological implications for the study of genre development. In N. Brügger (Ed.), *Web History* (pp. 257–277). Peter Lang.
- Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), 1–8. <https://doi.org/10.1057/s41599-024-02889->

- Lew, R., & Wolfer, S. (2024). CEFR vocabulary level as a predictor of user interest in English Wiktionary entries. *Humanities and Social Sciences Communications*, 11(1), 340. <https://doi.org/10.1057/s41599-024-02838-4>
- Nesi, H. (2024). Are we witnessing the death of dictionaries? *Ibérica*, 47, Article 47. <https://doi.org/10.17398/2340-2784.47.7>
- Preston-Kendal, D. (2025). Dictionaries in the web of Alexandria: On the dangerous fragility of digital publication. In G. Williams, M. L. Meur, & A. E. Peláez (Eds), *West Meets East: Papers in Historical Lexicography and Lexicology from Across the Globe* (pp. 31–41). Language Science Press. <https://doi.org/10.5281/zenodo.15394473>
- Rees, G. P. (2021). Discipline-specific academic phraseology: Corpus evidence and potential applications. In M. Charles & A. Frankenberg-Garcia (Eds), *Corpora in ESP/EAP writing instruction: Preparation, Exploitation, Analysis* (pp. 32–54). Routledge.
- Rees, G. P. (2022). Using corpora to write dictionaries. In A. O’Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (Second edition, pp. 387–404). Routledge.
- Rees, G. P. (2023). Online dictionaries and accessibility for people with visual Impairments. *International Journal of Lexicography*, 36(2), 107–132. <https://doi.org/10.1093/ijl/ecac021>
- Rees, G. P. (2024). Academic word families in online English dictionaries. *Lexikos*, 34, 437–468. <https://doi.org/10.5788/34-1-1947>
- Rees, G. P. (2025a). Exploring which aspects of an online monolingual learners’ dictionary can be investigated using the archived web. *International Journal of Lexicography*, ecaf020. <https://doi.org/10.1093/ijl/ecaf020>
- Rees, G. P. (2025b). Multimedia in dictionaries. In H. Nesi & P. Milin (Eds), *International Encyclopedia of Language and Linguistics* (3rd edn). Elsevier. <https://doi.org/10.1016/B978-0-323-95504-1.00685-2>
- Rees, G. P. (2025c). Making dictionary content accessible for people with visual Impairments. *Lexikos*, 35(2), 165–184. <https://doi.org/10.5788/35-2-2077>
- Rees, G. P., & Frankenberg-Garcia, A. (2025a). Writing aids for DDL. In L. McCallum & D. Tafazoli (Eds.), *The Palgrave Encyclopedia of Computer-Assisted Language Learning* (pp. 1–6). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-51447-0_55-1
- Rees, G. P., & Frankenberg-Garcia, A. (2025b). Dictionaries embedded: writing assistants and other tools. In *Reference Module in Social Sciences*. Elsevier. <https://doi.org/10.1016/B978-0-323-95504-1.00876-0>
- Roberts, J. C., Butcher, P. W. S., Lew, R., Rees, G. P., Sharma, N., & Frankenberg-Garcia, A. (2020). Visualising Collocation for Close Writing. In A. Kerren, C. Garth, & G. E. Marai (Eds.), *EuroVis 2020—Short Papers* (pp. 181–185). The Eurographics Association. <https://doi.org/10.2312/evs.20201069>
- Wolfer, S., & Lew, R. (2025). Supplementing CEFR-graded vocabulary lists for

language learners by leveraging information on dictionary views, corpus frequency, part-of-speech, and polysemy. *Humanities and Social Sciences Communications*, 12(1), 1151. <https://doi.org/10.1057/s41599-025-05446-y>

Yamada, S., & Xu, H. (2024). Learner's dictionaries. In E. Finegan & M. Adams (Eds), *The Cambridge Handbook of the Dictionary* (pp. 109–130). Cambridge University Press. <https://doi.org/10.1017/9781108864435.007>

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>

