Chapter 3

Dictionaries in the web of Alexandria: On the dangerous fragility of digital publication

Daphne Preston-Kendal
Humboldt-Universität zu Berlin

Digital publication is an attractive means of making dictionaries available to the general public, as has been widely documented by lexicographers and dictionary readers alike since the beginning of the "digital revolution" in dictionary publication over thirty years ago. However, commonly adopted means of publishing dictionaries pose dangers to scholarly integrity and to the long-term survival of lexicographical works. Adoption of practices used for the publication of some other online reference works and academic papers, indeed mirroring the survival characteristics of paper dictionaries, can help to mitigate these issues.

1 Models of digital reference publishing

The most common model adopted when publishing dictionaries online today is one which I have called the Oxford Model, after the *Oxford English Dictionary* Online which was probably the first to adopt it in full at its launch in 2000.[1]

[1] To be clear, I do not blame the OED editorial team nor Oxford University Press for adopting this model, though I am about to criticize it heavily. The digitization of the OED in the 1980s and development through,the 1990s into the online dictionary which launched in 2000 was, in every way, a pioneering process. The OED website first went online during the comparatively early days of the World Wide Web, when it was not yet clear what the implications and dangers of this model would be.

This is also not to say I consider the Oxford University Press entirely responsible stewards of the OED in this regard. However, 23 years later, a shift to a more sustainable model should

Daphne Preston-Kendal. 2025. Dictionaries in the web of Alexandria: On the dangerous fragility of digital publication. In Geoffrey Williams, Mathilde Le Meur & Andrés Echavarría Peláez (eds.), *West meets East: Papers in historical lexicography and lexicology from across the globe*, 31–41. Berlin: Language Science Press. DOI: 10.5281/zenodo.15394473

In the Oxford Model, nobody can purchase their own complete copy of the dictionary to keep on their own computer. Instead, a "subscription" is offered for sale, allowing the subscriber to view individual dictionary entries over the Internet. These entries are loaded one at at time from the one and only complete copy of the dictionary, which exists only on the servers of the publisher.

Because there is only one complete copy of the dictionary held on the publisher's servers, and because readers as subscribers can only load entries as they currently appear in that copy, readers have no guarantee that the entry they load today will have the same content as when they viewed it previously, or whether it might change (again) before they view it again in the future. The ability to publish the newest research immediately and perform ongoing revision of a dictionary is of course a great benefit of digital publication, but under the Oxford Model, readers have no way to go back and see exactly how an entry has been revised and when.

As subscribers to the dictionary website, rather than owners of a complete copy, readers are also beholden to the dictionary publisher on an ongoing basis. If they stop paying their subscription fee, they lose the ability to read the dictionary and cannot even see entries they have previously looked at, unless they had the foresight to save them to their local disk. On the other hand, if the publisher were to go out of business, decide that hosting the website was too unprofitable, or otherwise lose interest, they may simply take the website offline, at which point nobody can read the dictionary any more.

The Oxford Model in its complete form is the most irresponsible way to publish dictionaries online from the point of view of ensuring long-term, sustainable accessibility. The reasons for this, if they were not immediately clear from this description, will be explored in more detail below. Despite this, the Oxford Model is also very widespread: in the German national database of electronic academic resources, the *Datenbank-Infosystem*,[2] 1,373 resources described by "dictionary, encyclopedia, reference work" (*Wörterbuch, Enzyklopädie, Nachschlagewerk*) are accessed over the web in exchange for a licence fee as of writing.

There are a few variants of the Oxford Model in use by some dictionaries, each of which tempers the problems with it. Some online dictionaries are published under an Open Access Oxford Model, in which there is no subscription fee, but the publication in other respects follows the system outlined above. Others have

certainly have been made long ago. The Oxford University Press has declined requests from scholars for access to digital copies of older OED editions. Decision-making with regards to the OED in recent years has appeared to be increasingly profit-motivated, to the detriment of the long-term scholarly value of the work.

[2] https://dbis.ur.de/

adopted an Oxford Model Without Revisions, where there is at least an informal policy that entries are not revised after they have been published (except perhaps to fix typographical errors and other, very minor details, of the same variety of change which may also be made between printings of the same edition of a paper dictionary). While both of these variants maintain some problematic aspects of the Oxford Model, they do at least mitigate some of the others. How and why will also be explored below.

2 The dangers of centralization

A key aspect of the paper model of publication is the distribution of complete copies for a one-time cost borne by the purchaser. While this may seem obvious, it is also notably different from the Oxford Model described above, to the extent that proposals to reinstate this crucial aspect within the context of digital publication can seem radical by comparison.

Why, then, is the distribution of dictionaries from a centralized authority only so problematic? Some problems of models in which the only complete copy of a text is held by the publisher are outlined above. The technological structure of the World Wide Web essentially encourages this model for the publication of all kinds of media online, but this situation is bad for the long-term accessibility of that media. New media researcher Bret Victor has informally called this situation the "Web of Alexandria", in reference to the burning of the library of Alexandria, a disastrous loss of knowledge caused by the failure to make and distribute duplicate copies of works which were instead held in one central location (Victor 2015). While physical natural disasters or human accidents such as file deletion remain a danger to digital texts stored on the servers of a single institution, this model is insidious in other ways.

One has to do with how one of the advantages of digital publication – the ability for scholarly dictionaries to be continually revised to reflect the latest research – is applied in a centralized system. It has become quite normal for entries to be revised in their online editions. A description of how this occurs in practice under the Oxford Model is described in Simpson (2012): the OED1 and OED2 sub-entry for *market garden* was upgraded to a full entry in 2000, the first year of OED3 entry publication. At the time, the oldest quotation was from 1811, an improvement on the 1840 date given in the older edition. Two subsequent revisions pushed this back first to 1793, then to 1727. However, these revisions are "invisible" to the reader, and there is no way to trace when these antedatings were added. Indeed, there is no way to trace the specific dates of any changes to OED3 articles after their initial publication, nor what they looked like at a particular previous point in time.[3] It is entirely conceivable that a study of, for example, word coinages of the early 19th century might have cited this OED3 entry at the point in time when it claimed the word was coined in 1811. Future readers of that study following this claim back to the source are left to wonder why its author missed or disregarded the earlier uses since added. This alone undermines basic scholarly practice.

Market forces and changing institutional interests can also pose a danger to the long-term accessibility of dictionaries under the centralized digital publication model. If a publisher decides a dictionary project is too unprofitable or loses interest, it may simply decide to remove the dictionary from the web entirely, so it becomes effectively entirely lost to history. We have already witnessed cases where publishers have made decisions like these: the tragic case of the *Historical Dictionary of American Slang* (henceforth HDAL)(Lighter 1994, 1997), cancelled by its publisher after only two of its planned four volumes, is well known. In a paper publication model, the completed dictionary parts are at least accessible by those who have already purchased copies; in an Oxford Model online environment, this kind of commercially-motivated decision would simply result in the total inaccessibility of the work already done. In fact, we have even seen a very similar dictionary become a victim of exactly this failure mode of the Oxford Model: *Partridge Slang Online*, an online edition of the *New Partridge Dictionary of Slang and Unconventional English* (Dalzell & Victor 2005), disappeared without trace from the web after about three years without any apparent explanation.[4] Fortunately in this case, the dictionary had also previously been published on paper, so the dictionary text remains accessible in libraries and on private bookshelves around the world, albeit without updates, advanced search tools beyond alphabetical headword lookup, and other benefits of digital publication.

Finally, while the ability to update dictionary entries on an ongoing basis is certainly a benefit of digital publication, the practice of doing this in a centralized

[3] Recently the OED3 website has added an indication of the month and year in which each entry was last revised, but there remains no indication of what was actually changed at each revision, nor any way to actually go back to an older version. This indication also often appears to be inaccurately forward-dated: as of writing, the entry for *market garden* claims the entry was last revised in March 2022: but there are no textual differences compared to the version of 2012 which appears in Simpson's paper.

[4] *Partridge Slang Online* was online at http://www.partridgeslangonline.com/ from approximately January 2013 to September 2016 according to the Internet Archive WayBack Machine, which has archived the home page, front matter, and a small number of sample entries, but not the dictionary as a whole. After September 2016: only error pages are archived.

publication model makes the dictionary vulnerable to political attacks. Numerous dictionary projects, even those of a scholarly nature, have been the victims of political interference typically by totalitarian regimes,[5] but those regimes could not retroactively alter the contents of volumes of dictionaries already published (except by official edicts to libraries to censor their copies, which is reported to have happened in some cases). With centralized digital publication, this kind of change to entries for political reasons can be carried out immediately, affecting all users, and potentially without any trace of the older versions remaining.

None of this, however, need dissuade us from digital publication and its benefits entirely. Instead, an alternative model is required which retains these advantages over print while providing the long-term stability which was intrinsic to printed copies. As we will see, the low cost of digital publication can actually even lead to better and more widespread long-term accessibility than is possible with paper, provided the right technical and administrative choices are made about how the dictionary is actually published.

3 An alternative model

While digital dictionaries, especially scholarly dictionaries, are unique among digital texts in the extent and importance of their microstructure, lexicographers can still hope to learn from the experiences of other kinds of scholarly resources in exploring the possible models of digital publication.

A particularly instructive model, one which I hold to be the gold standard for digital publication of scholarly works, is that adopted by the *Stanford Encyclopedia of Philosophy* (SEP, Zalta 2006; see also Hammer & Zalta 1997, Allen et al. 2002). In the Stanford Model, an endowment has been established under a membership model which universities, libraries, and other institutions may join for a (significant, but one-time) fee. In exchange for this fee, the institutions obtain the right (and implicitly the responsiblity) to archive the SEP on their own computers in perpetuity; if the SEP project ever closes down or becomes otherwise unavailable, they then gain the further right to re-host the entire encyclopedia on a public website to ensure general access to the text remains available. The encyclopedia is revised in versions which carry the dates of their publication explicitly, and all older versions of the encyclopedia and of the individual articles remain available in archived editions, both on the official website and in the

[5] As examples, see for descriptions of the ideological pressures exercised on lexicographers under the two dictatorial regimes in Germany in the 20th century, (Lea 2009) and (Zielinski 2010).

archived versions held by member institutions. Finally, because the membership scheme supports an endowment and only the returns on the investment are used to actually support the project, Stanford promises to return the membership fee with interest to members should the SEP project ever shut down. As of writing, over 500 institutions worldwide have become members of this scheme.[6]

This model in its entirety is incredibly ambitious. However, I will outline the most important aspects which can be applied to digital lexicography, dealing first with what might be termed the technical aspects, then an overview of the financial aspects.

The first technical aspect, as implied by the previous section, is that each member institution has under its control complete copies of the encyclopedia. They thus retain access to it, no matter what should happen to the producers and publishers. As applied to digital scholarly dictionaries, however, it is important not only that the text itself be made available, but that it be in machine-readable form; that is, with the microstructure encoded in a database file usually using descriptive XML or SGML markup. It is this machine-readable database structure which enables the advanced search capabilities which users of digital scholarly dictionaries have come to expect and rely upon. With machine-readable database files, this functionality can be made to work again long after the original software has become technologically obsolete.

The second is that this archival capability is, de facto, not strictly limited to member institutions; rather, because the SEP is available as open access in web page form without any authentication of users, the web pages themselves are picked up and archived by web crawlers such as the Internet Archive WayBack Machine and potentially by other private "webpage capture" services such as archive.is. These static web page archives can only be considered machine readable in a limited sense, and are not ideal for the use cases for machine-readable text previously outlined – but simply making an existing dictionary website open access represents an important and easy first step to allow archiving by third parties.

The third technical aspect is that not only do older versions happen to be accrued by member institutions over the period of their membership, continued distribution of older versions of the text is actually explicitly foreseen by the model. The ability to continue to distribute older editions and thus maintain the value of scholarly citations (as mentioned above) is an advantage of digital publication which has, in lexicography, remained thus far comparatively unexplored. It is another consequence of the fact that digital reproduction is close to free,

[6] https://plato.stanford.edu/support/commitments.html

compared to the unattractive paper and ink costs associated with continually making older editions of a printed dictionary available.

These first two of these three technical aspects are not unique to the SEP but are in widespread use for the distribution of academic papers under the open access model.

I will now briefly discuss the financial aspects of the Stanford Model, though with an acknowledgement that the specific funding situations of individual lexicography projects differ significantly from one another.

Notably, the Stanford Model as applied for the SEP is concerned only with covering the costs of publication of the encyclopedia, not of paying the authors of its articles, which they write on a pro bono basis. Most lexicography projects involve a small integrated team, rather than a large distributed worldwide team of experts each writing articles in their own field of specialization – this difference necessarily implies that the SEP model, whereby authors consider their work on the dictionary to be part of their output of academic writing in their careers, is not workable for lexicography. While most European historical dictionary projects are funded as research projects under the auspices of a university or academy of sciences, and the costs of producing the text thus subsidized away for the eventual publisher who must only cover typesetting and printing costs, the OED and some other dictionaries of potential scholarly relevance (such as the "Unabridged" dictionary published by the Merriam-Webster company in the US) are commercial undertakings. This, indeed, was the case with HDAS, where it was likely not the printing but the editorial costs which motivated the early cancellation of the project.

It is difficult to know what to suggest in such cases, other than that the preparation of scholarly dictionaries in the Anglosphere, where currently carried out for commercial or quasi-commercial publishers, should move towards the European model. The association between the OED and the Oxford University Press goes back to the first edition, but perhaps it is time for this arrangement to be reconsidered, and for the editorial work of the OED to come under the aegis of and be funded by, for example, the University of Oxford's Faculty of English, rather than its University Press. This is merely an idle suggestion, however, and made in the acknowledgement that European scholarly dictionary projects have historically moved much slower even than English ones, probably largely because of this model in which lexicographers are also employed as academics with their own teaching and research schedules, rather than as full-time dictionary writers.

4 Further work

It was suggested at the conference that some clear guidelines should be prepared for the sustainable publication of digital dictionaries. If lexicographers, lexicographical historians, and dictionary publishers feel such guidelines would be useful, an expert group could be assembled to use the experiences of lexicographers working on projects of various kinds, as well as those of librarians and archivists, to produce a thorough but realistic set of recommendations. Since scholarly dictionaries in particular are, in the digital age, increasingly being published "in house" by the department of the university or other institution which produces them, leaving decisions about publication mostly in the hands of lexicographers, such guidelines may be valuable especially for smaller teams which may lack specific expertise in this area. Several such smaller projects were presented at ICHLL12.

Nonetheless, I hope this paper's recommendations – that wide distribution of machine-readable dictionary texts in their entirety and with their complete revision history should be made an integral part of the publication of lexicographical projects – offer a useful starting point until such guidelines are published.

The existing infrastructure for open access journal and data-set distribution provides a useful starting point for the technical means of achieving this. Another project of Stanford University, LOCKSS (Reich & Rosenthal 2000),[7] provides open-source software used by a number of "preservation networks" which automatically create copies of digital resources in linked member institutions around the globe. One network established by Stanford currently archives 52 million journal articles and nearly half a million ebooks; in addition, archives established by the governments of the United States and of Canada distribute and save those countries' digital national archives, among other networks which preserve data for the future which cannot yet be made available to the public. LOCKSS is clearly a proven technology for this purpose, but given the unique requirements of digital dictionaries as structured texts, it would perhaps make most sense to establish a new network using this software which all historical lexicography projects could use to share their published work.

Until sustainable electronic dictionary publication as the norm becomes a reality, stopgap solutions must be found. One simple such stopgap is to continue to publish and buy dictionaries in paper form. This can be achieved through pressure on publishers and libraries to keep print alive even if digital publication is still perceived as the superior option. The paper publication model, as

[7] https://lockss.org/

outlined above, has intrinsically good survival characteristics: dictionaries such as *Partridge Slang Online* which have already died in digital form survive in paper form. “Retro-redigitalization” from a paper source of a dictionary which was originally made digitally, but whose digital form has been lost, may be a realistic last-resort possibility to recover a digital text.

In the era of online journal publication, it has become the norm for (sometimes extensive) supplementary material to be made available with the electronic version of an article. As a further stopgap measure, as long as Oxford Model reference resources exist, I propose that it should be made the norm that these supplementary materials should include PDF archival copies of the complete versions of any dictionary entries cited, when those entries were accessed from an Oxford Model publication. Given the low cost of electronic storage and distribution, this is technically plausible even for articles which cite very many dictionary entries, though in this case legal difficulties may arise due to copyright law. However, this practice alone should be sufficient to guard against the problem of being unable to trace back citations after an Oxford Model dictionary has been revised to obscure an original reference.

A further consideration related to the long-term retraceability of dictionary citations is the long-term reproduceability of research conducted using text search tools on online dictionaries. This implies that the exact search algorithms used need to somehow be reproducible. Even apparently simple algorithms can have very complex requirements in reality: a case-insensitive search is actually intrinsically language sensitive (as one example, ⟨i⟩ is not the lower case form of ⟨I⟩ in Turkish, unlike in most languages). This alone makes the task of searching an etymological dictionary, which typically mentions words from dozens of different languages in its text, much more complex. This complexity affects the reproducibility of research based on text searches, because the approaches used by a particular search software must be fairly exactly replicated in order to obtain the same results. The simplest approach would be to make all dictionary search software open source so that the same software can be used by everyone on their own computers. Code, though, intrinsically has worse survival characteristics than data: software platforms are constantly evolving, but data storage standards such as XML are essentially permanent. Making software which can still be used many years in the future is a broader problem in computing research, but the number of complicating factors in the design of linguistic applications makes the issue even trickier for our use cases.

5 Summary

My purpose in this article is certainly not to dampen the great enthusiasm for digital publication which lexicography has shown over the last few decades. On the contrary, I welcome digital publication and recognize the great power it offers – much of it as-yet still totally unexplored by lexicographers. I instead wish to heed caution and encourage more careful thought about the precise means by which digital dictionaries as texts are made available to the public.

Fixing this problem will require changes in the technical means of dictionary publication, as well as potentially in how scholarly dictionary writing is financed. Some of the technical problems involved have already been solved by other digital publications facing similar issues; others will require active research to resolve. The financial problems, where they exist, will likely have to be resolved by individual dictionary projects, since sources of funding already differ greatly between projects. Above all, solving this problem requires a recognition by dictionary sponsors that a scholarly dictionary is not a source of profit, but a contribution to human knowledge, and must be treated as such.

Indeed, solving this problem is a necessity if dictionary making as a scholarly pursuit is to continue, because matters of basic scholarly integrity are at stake. We must save our dictionaries from the Web of Alexandria.

References

Allen, Colin, Uri Nodelman & Edward N. Zalta. 2002. The Stanford Encyclopedia of Philosophy: A developed dynamic reference work. *Metaphilosophy* 33(1). 210–228. DOI: 10.1111/1467-9973.00225.

Dalzell, Tom & Terry Victor (eds.). 2005. *The new partridge dictionary of slang and unconventional English*. London: Routledge.

Hammer, Eric M. & Edward N. Zalta. 1997. A solution to the problem of updating encyclopedias. *Computers and the Humanities* 31(1). 47–60. DOI: 10.1023/A:1000418920193.

Lea, Henry A. 2009. Dictionary-making in the Third Reich: The case of *Trübners Deutsches Wörterbuch*. *Seminar: A Journal of Germanic Studies* 45(4). 369–386. DOI: 10.3138/seminar.45.4.369.

Lighter, Jonathan E. 1994. *Random House historical dictionary of American slang*, vol. 1: A–G. New York: Random House.

Lighter, Jonathan E. 1997. *Random House historical dictionary of American slang*, vol. 2: H–O. New York: Random House.

Reich, Vicky & David S. H. Rosenthal. 2000. LOCKSS: Lots of copies keep stuff safe. *New Review of Academic Librarianship* 6. 155–161. DOI: 10.1080/13614530009516806.

Simpson, John. 2012. What has the OED become? *Nordiske studier i Leksikografi* 11. https://tidsskrift.dk/nsil/article/view/19131.

Victor, Bret. 2015. *The Web of Alexandria.* http://worrydream.com/TheWebOfAlexandria (14 May, 2023).

Zalta, Edward N. 2006. The Stanford Encyclopedia of Philosophy: A university/library partnership in support of scholarly communication and open access. *College & Research Libraries News* 67(8). 502–504. DOI: 10.5860/crln.67.8.7670.

Zielinski, Lech. 2010. Ideologie und Lexikographie: Die Ideologisierung des *Wörterbuchs der deutschen Gegenwartssprache* von Ruth Klappenbach und Wolfgang Steinitz. In *Danziger Beiträge zur Germanistik*, vol. 31. Frankfurt am Main: Peter Lang.