Pasadena: A New Editing System for the Oxford English Dictionary

Laura Elliott and Sarah Williams
Oxford English Dictionary,
Oxford University Press,
Great Clarendon Street,
Oxford OX2 6DP

Abstract
The Oxford English Dictionary and the software company IDM together designed a new editing system that went live in June 2005. The system is called Pasadena. Though specially designed for OED, the system incorporates software which is commercially available for other dictionary projects. This paper explains some of the dèsign principles behind the Pasadena system and reviews the performance of the system after six months in live operation.

1 Why a new system was needed

The *Oxford English Dictionary* (OED) was digitised in the mid-1980s to allow the creation of new print and electronic editions.[1] The digitisation project included the development of software which was then successfully adapted to make an editing system ready for the start of the comprehensive revision of OED in 1993.[2] That system supported OED editors through the launch of *OED Online* in 2000, but eventually it started to show its age: it had not been conceived for the creation of regular updates to a web publication; it did not cover all the editorial activities involved in OED revision, and subsidiary systems were growing up round it – always a sign that something different is needed. And it was becoming hard to support technically: for example, it had to run on hardware which was incompatible with Oxford University Press's email and administrative systems. By 2003, it had become urgent to replace the existing system, and by this time, it was recognised by OUP that this would be an excellent opportunity to review the editorial system's purpose and scope.

2 Lexicographical involvement with the system design

The style of the project was always to maximise the participation of lexicographers in the design of the system. For this reason, it was an enormous bonus for potential software devel-

[1] See http://www.ariadne.ac.uk/issue24/oed-tech/ 'How the *Oxford English Dictionary* went online', Laura Elliott, 2000.

[2] See http://www.oed.com/about/oed3-preface/ John Simpson, 2000.

opment partners if they could demonstrate a good knowledge of lexicographical use of computer systems: it would have been overwhelming to have to teach a supplier to understand this on top of everything else involved. Formal user requirements were documented early in 2003 and a tender document for the system was distributed to a shortlist of software developers. The French firm IDM demonstrated the kind of knowledge of computerised dictionary-making that we were seeking, and they were chosen to undertake a "blueprint" design phase. From July to October 2003, a joint team from IDM and OED worked together on a functional specification, on the basis of which IDM could quote a confirmed price for the development of the system. This confirmed figure was approved by OUP in November 2003, after which OED and IDM starting working together on an extremely detailed specification and design for another nine months. A partial beta version of the system was ready in September 2004, and work continued on refining this till the final delivery of the live system in June 2005. Throughout this time, OED had a number of lexicographers working part-time on this project.

3 Design issues for Pasadena

3.1 Design principles

The overall aim was to provide a system that automates what can be automated, and leaves the lexicographer to do lexicography, which should be true of any adequate dictionary editing system. The system should also have a degree of built-in flexibility for changes in working practice, as these are bound to occur on such a long-term editorial project. Especially in the case of the OED's old editing system, there was also a pressing need to integrate in one system the many different activities involved in making the dictionary.

To describe the problem: for good reasons at the time, the 1993 editing system had been restricted to editing dictionary entries one by one, and to searching, separately, work-in-progress and the various kinds of ancillary electronic material collected over by OED over fifteen years. The mark-up underlying the OED's electronic text still reflected the philosophy of the original digitisation project, designed to retain every feature of the original print publication and in an idiosyncratic mark-up style. Quotations collected electronically by the OED's reading programmes were kept in separate databases with separate editing systems. Administrative workflow systems had been established for various types of work chasing and monitoring, but these were operating separately from the main editorial system. Automated validation of the text was limited. The text was full of comments, administrative and editorial, which were useful in situ, but needed to be searched round or stripped out for some purposes. The OED's bibliography was scarcely computerised at all. It was definitely time to look at the possibilities for a more integrated approach.

3.2 Scope of the redesigned system

There is only space in this paper to sketch the practical implications of the resulting complete redesign of the system.

3.2.1 Mark-up redesign

The mark-up of the dictionary was converted to XML, the standard electronic format of Oxford University Press (OUP). The basic principles of the XML mark-up design were that each editorially significant section of the text should be "boxed" by tags, and within that "box", spacing and punctuation and standard text would be stripped out of the underlying data, to be shown on screen or in print whenever necessary. The intention was that a computer program should reliably be able to identify elements of the text, and that lexicographers would benefit, despite heavier tagging than they were used to, from the excellent text presentation and manipulation permitted by this clear-cut mark-up. For example, it was only through this new mark-up that automated chronological and alphabetical sorting was made possible for quotations, variant word forms and so on.

3.2.2 Cross-references

OED at the start of revision in 1993 contained roughly 600,000 cross-references, but the electronic text at that point contained no active links between cross-reference source and target. Data conversion rules were applied during the preparation of data for Pasadena which linked over 80% of source items and their targets. These active links could then be used by cross-reference checking software. Cross-references in Pasadena also update automatically when the relevant sense number or entry homograph number is changed.

3.2.3 Handling external research requests

Research requests (from lexicographers to consultants and researchers anywhere in the world) used to be managed just with email and tracking comments embedded in entries. OED can generate as many as 400 such requests a week so these methods were very time-consuming and error-prone. It was also a problem that external researchers had no access to up-to-date work-in-progress, so their replies could be out of step with in-house work.

IDM constructed a work tracking system for these requests using the web-based package Oracle Workflow. The package enables administrators to define workflow for any particular type of request, and to set up automatic alerts to tell a lexicographer when answers arrive, or a researcher when requests are running late. It is possible within the Pasadena interface to attach a request to a particular piece of data (so the request remains valid even if the item is moved within or between entries), to make or review research requests while editing an entry, and to move directly to the relevant point in the current version of an entry from the workflow software. Staff researchers have been trained to use the new system in full, but consultants can use the web-based look-up facility to see OED work-in-progress without needing to learn to use the entire research management system.

3.2.4 Integrating workflow with the editing system

Under the old system, progress was generally measured in senses revised or drafted ready for the next level of editorial review, but there was no electronic help with counting these senses. Microsoft's Project Central was used to record schedules and for weekly editorial

progress reporting, but this software did not interact at all with the old editorial system. Reports could not be derived from Project Central on group progress through work batches, as opposed to group progress per week, so work groups devised their own individual progress reports. Sizing batches of work required a considerable amount of arithmetic with search results.

There were problems in knowing exactly what had been done to an entry at a given stage in the long journey from initial revision or drafting, through specialist review, to approval for online publication. Editors were clear what their tasks were at a given editorial stage, but they depended on a complex mixture of internal evidence, embedded comments, intranet lists and paperwork, to confirm whether all the tasks of a previous stage had been completed as expected.

It was certain that requirements for work scheduling and reporting would change over time, for example when senior managers changed. So any computerised help with the first two problems must also allow for reconfiguration if the management context altered.

The solution that OED and IDM developed was this. Pasadena has an administrative interface to its database of dictionary entries showing how many senses they contain, of various different types, and provides a tool for managers to create work batches of the right size. A managing editor can define and redefine workflows for different editorial activities, each made of a sequence of stages. Each stage has a number of optional tasks. A list of entries can then be associated with a particular work group, a particular lexicographer, and a particular workflow. When a lexicographer works on an entry in such a list, the editing interface displays the tasks to be done at this stage. The lexicographer has to record each task as completed or postponed, and then when all tasks have been acknowledged, can "sign off" at that stage. As a result of recording this information, any other lexicographer can look up a summary of what has been done to this entry at previous editorial stages. From this information, the system also derives up-to-date progress reports with sense counts. This information can be exported to Project Central to help with any necessary rescheduling; it was not decided not to drop Project Central completely because its recalculation function for time and resources is useful and not easily replicated.

3.2.5 Bibliographical standardisation and quotation management

A long-term task of OED revision is to verify the sources from which quotations are drawn. This may involve correcting both the content of quotations and the citation details, and these corrections may change the chronology of the senses within an entry. For efficiency, this task is primarily undertaken across the alphabet by author.

The old system only allowed re-verified citations to be corrected one by one – for example, if a change to the spelling of the name of the First Quarto of a Shakespeare play was required, every single quotation from that edition would have to be corrected separately. We needed to find a way to do such a correction to a set of quotations, but not blind, to allow for any exceptional cases.

With IDM, OED investigated the scope for creating a database of authors and their works to which individual quotations would be linked and from which citation details would be

generated in their correct form. It was established that this approach would add considerably to the complexity of the system without much benefit because so much would still need to be disambiguated editorially. What was agreed instead was that the quotation content and its location details would be held linked to the citation reference, so that a single citation reference could be corrected once for all related quotations. But the interface allows for quotations to be excluded from the change where necessary. Rather than including authors and sources in the main database, web pages have been constructed containing information on authors and sources, arranged pragmatically to hold all the information collected by OED to help with correct citation and bibliographical verification (for example, authors' birth and death dates, on the publishing history of frequently cited sources, on authors with complex name changes and involvement in shared authorship). IDM made it possible for these pages to be linked to the quotations from each source, and provided an interface from which bibliographers can easily change the link between quotations and sources where research or corrections make this necessary, so that the inter-linked information pages can be kept up to date. In addition, editors can copy an existing citation style with its link to its source, useful when adding, say, additional quotation from an approved edition of Shakespeare.

Quotations that occur in dictionary entries are displayed with their citation details when dictionary entries are displayed or edited. The dictionary entry as it is held in the database holds only the link to the quotations database.

Quotations collected through OED's several directed reading programmes are included in the same quotation database, though they may not yet be linked to any specific dictionary entry.

This software has added a dimension to the OED editing system which is enormously powerful, and has allowed much swifter correction of problems with references, though it has not yet solved all problems of scale: sometimes the system struggles with the large number of related quotations. The bibliographers' interface is an embryonic electronic annotated bibliography for OED of considerable elegance and breadth.

3.2.6 Search engine

Pasadena uses IDM's XML search engine, skxml, which can search across any XML material with a powerful range of wild-carding and optional settings for case, accentuation, inflections, searching at different levels in the tag hierarchy and so on. The search engine can find any Unicode special character. For the first time, an editor could search for something both in OED and in all its ancillary material in a single search.

3.2.7 Practical help with text manipulation

OED editors spend a considerable amount of time reorganising and adding material to existing dictionary entries. A few examples of improvements in efficiency in doing this follow. Pasadena provides automated help with sorting and reordering various elements of the text, with scope to add or change sorting routines. Block moves of text have been made easier. Content validation, based on configuration scripts completely under OUP's control, can be invoked for whole entries or parts of them, and can be refined ad infinitum. The embedded

annotations that used to pepper the text (one long entry used to contain nearly 2000 of these) can now be viewed, added or deleted in a side panel that keeps the main text clean.

4 Review after six month's use

4.1 Editorial acceptance

The Pasadena system went live on 15 June 2005. The go-live itself was extremely smooth, and over the following six months the system has continued to be very stable, particularly given the complexity of its architecture, and the number of different applications involved. Happily, early concerns that lexicographers would find it difficult to adjust to the new mark-up have not been realized. Similarly, design decisions to separate workflow information from dictionary data, and to remove bibliographic citations from the Entry Editor, were quickly accepted by editors. Almost all members of the department were able to resume their editorial work in the new system after a brief training period. Some who had expressed scepticism at the beginning of training were later keen to state how impressed they were by the capabilities of the new system. The previous best editing rate using the old system was reached within three months of using the new system, so the editors are now moving towards more ambitious targets.

4.2 Improvements needed

It was an important part of the design of the system that as much as possible should be configurable once the system was live without altering the underlying software, partly to allow for changes in editorial policy and working practice, partly to allow for the easiest possible adjustments to correct specifications that turned out to be problematic in practice. The reconfigurations needed over the first six months have principally been in the areas of mark-up and workflow.

The system allows the underlying XML DTD to be adjusted, except in relation to a very few items of data which are significant to underlying software. Some changes have already been made to the DTD, to correct problems with the simplification process currently in place for online publication, and to accommodate difficult or exceptional cases in the data. More of these latter changes will inevitably be necessary in a text as complex as OED. These changes are relatively straightforward to make, and have not been too disruptive to ordinary editors.

It is expected that in due course editors will review the style of mark-up and decide whether it needs to be made more editor-friendly. There is a natural tension between a style of text encoding that makes a text easily machine-readable, and a style that reflects a non-technical editor's view of the text, and some of the structural mark-up intended to help with machine-readability, and to ease operations like the upgrading from compounds to main entries, is in fact too cryptic for editors to use with comfort. Again, at least the fact that the DTD can be changed, and global changes made relatively easily to the text, makes it possible to contemplate such a review.

Slightly more problematic has been the process of refining the configuration of workflow processes to meet practical editorial requirements; this latter has been allied to difficulties in extracting the precise information needed for reporting purposes from the Schedule Manager

interface. As far as the latter is concerned, it has been agreed that an enhancement is needed to the scheduling and reporting tool, to allow managers to create bespoke reports. This is likely to use extant software, integrated into the Pasadena interface; it is a relatively small system modification which will be of immense benefit to editors. Alterations to the processes are being planned, and to the configuration of the so-called "characteristics" which identify sub-units of an entry to be counted in workflow; the need for these system features, and the form they need to take, are becoming clearer with increased experience of using the new system.

Despite all the configurability that is allowing Pasadena to be tuned while in use, there remains an area in which fundamental adjustments to the system are likely to be needed. The problem is with the handling of very large, or very extensively linked, units of data. At present very large entries are difficult to edit and very slow to save in the Entry Editor.

A comparable problem is the difficulty of editing bibliographical citations (i.e. references) to which a large number of quotations are attached. This latter impedes the bibliographers' efforts to normalize the data, which was one of the major anticipated benefits of Pasadena. It is therefore a very high priority in the re-engineering of the system. Unlike the configuration changes mentioned above, these are problems to which the solutions are not obvious, nor pre-empted by the system design.

This is the area in which the unusual structure and size of the Oxford English Dictionary has posed the greatest difficulty to the developers, despite their careful assessment of the technical risks and consultation with experts; such challenges are likely to remain, at least to some extent, for the lifetime of the system.

4.3 Benefits realised already

Sometimes, at the start of using a system planned to last for a number of years, the need for further improvements weighs more heavily on the users than the benefits immediately won. Fortunately for Pasadena, there have been a number of straightforward successes. The two main areas to have benefited conspicuously already from the new system are library work, and the practical business of editing. The automation of the administration of library work has been highly successful. In the first week after go-live, one research administrator was reporting that her workload was 15% of what it had been. For editors, although aspects of the system remain characterized as 'clunky', it is clear that the anticipated improvements to the experience of editing have been realized to a very considerable extent. Sections of text, such as sense units, are much easier to manipulate in the new system, and this has clearly been of value, as has the related automatic renumbering function. The ability to validate content within the entry itself, rather than from error reports, has been a notable step forward. And the visibility of what has happened to an entry, which is provided by the detailed work tracking information, is also of great value.

5 Conclusion

Earlier directors of the OED had wisely hoped to be able to invest in the evolution of a new editorial system over a number of years. Instead, the system eventually had to be re-

placed in one large project, with the risk that the scope of the project would be over-ambitious. OED have gone beyond replacing their editing tools to modernise their mark-up, introduce workflow, and now have a bibliographical database, and computerised administrative tools. Of course there are teething problems, but also great successes of design, foresight, and collaboration. This is an enormous step forward into freeing lexicographers to do lexicography, which should be the aim of every dictionary-making system.

References

A. Dictionaries

OED Online. Oxford University Press. http://dictionary.oed.com/

B. Other Literature

Elliott, L. 'How the *Oxford English Dictionary* went online' http://www.ariadne.ac.uk/issue24/oed-tech/

Simpson, J. (2000) http://www.oed.com/about/oed3-preface/