

Reprinted from *The Times* Friday May 9 2003

## A database for old Times' sake

By Jim McCue

THE full research potential of *The Times* is being unlocked, or perhaps unleashed. Complete pages and all the articles from our first 200 years will shortly be available online in a form that makes possible astonishingly sophisticated searches — although to begin with only in subscribing libraries.

Gale, part of Thomson Learning, is scanning all the issues published between 1785 and 1985 — nearly a million pages and some ten million articles — to create The Times Digital Archive. To that is added technology that can read the original and sometimes erratic typesetting and locate words or phrases across the centuries. As well as editorial matter, all display and lineage advertisements and even the crosswords are being scanned.

In just a few seconds from entering a search term, the user is shown statistics of the number of occurrences, by category: advertising, editorial and commentary, news, picture gallery, business, features and people.

Beneath this are chronologically ordered thumbnail images of each relevant page, with a red border around the article in which the search term occurs. The border shows where the article is on the page, and this and the headline indicate its likely nature. A click will then produce either a larger image of the entire page or an onscreen "cutting", reproducing the article in clear facsimile, which can then be printed.

The text of the entire two centuries — a run of papers that fills a 40-yard corridor with shelves three deep on both sides in the Wapping basement — should be available by September. At present the years 1880-1985 are online, so that *The Times's* coverage of most of the 20th century can be trawled.

Search, for instance, for the word "computer", and the program finds early examples from when it used to mean a person who does calculations. *The Times* first recorded its use to denote a machine in November 1946 in a report of an impressively perceptive speech by Lord Mountbatten, in which he referred to the "electronic numeral integrator and computer (Eniac), employing 18,000 valves".

Describing the coming "revolution of the mind", Mountbatten imagined a machine that "could even be made to play a rather mediocre game of chess!" And the report went on: "In the field of memory alone, it seemed likely that Man was to be provided with vastly greater and speedier access to the inherited knowledge of the ages than he was able to command at present." The two terabytes of images and text in this new database certainly fulfil that promise.

But although the automated reader successfully identifies words printed in italics or hyphenated across lines, it is far from infallible. It can throw up a lot of gubbins, as a search for a non-existent word shows. For instance, "boyb" is said to appear more than a hundred times in the available years. The program is misreading words such as "boys", mostly in the small type sizes of classified advertising. Inevitably, then, the success rate with other search terms will be less than 100 per cent. Furthermore, misprints in the original are not utterly unknown.

Nevertheless, the technology is awesome, and the results are generally better (and clearer) than the equivalent searchable version of *The New York Times*. The program can also do fuzzy searches, identifying words similar to a keyword — revealing, for instance, that in the 1920s *The Times* referred occasionally to "Herr Adolf Hittler".

Practically the only thing the program will not do — and this is a commercial, not a technical restriction — is convert the articles it locates to plain text that can be cut, pasted or e-mailed.

*Cont'd...*

# The Times Digital Archive

1785 – 1985



The Times Digital Archive is the largest project so far undertaken by Gale, a company that specialises in commercial provision of large historical archives. Its next is to digitise the entire text of virtually every 18th-century book printed in Britain — 150,000 titles.

Initially, The Times Digital Archive is being marketed to libraries (some 800 of which hold Gale's microfilm edition of *The Times*, from which the electronic version is largely being prepared). Online subscriptions start at £2,100. One must hope, however, that prices will fall in future as the initial costs are amortised, or that pay-per-view facilities will be made available. The sheer power and potential are too great to keep locked up.

One of the first subscribers was the Oxford English Dictionary, delighted to be able to search for earlier occurrences of words than those in its records. But every kind of history will be facilitated, including genealogy, biography, social history and the history of ideas.

Not only is the search facility unprecedentedly comprehensive, but it will also produce negative evidence, showing that something or someone was mentioned much more rarely than one would suppose, or not at all. It is easy to ascertain, to take a couple of literary instances, that Joyce's *Finnegans Wake* and Larkin's *The Less Deceived* were not reviewed on publication.

Such phenomenal tools will inevitably change the nature of research, as Google already has. There will be fewer opportunities to discover sources overlooked by everyone else, but all the more need to be imaginative in winnowing, making intelligent connections and deciding what evidence really matters, and what it signifies.

Meanwhile, for Times journalists there is a chastening sense of contributing day by day to a vast and enduring encyclopaedia.

*The supply of the material by The Publisher does not constitute or imply any endorsement or sponsorship of any product, service, company or organisation. Material may not be edited, altered, photocopied, electronically scanned or otherwise dealt in without the written permission of The Publisher. Times Newspaper, 1 Pennington Street, London E1 9XN tel: 0207 711 7888 email: [enquiries@nisyndication.com](mailto:enquiries@nisyndication.com)*

**For further information on *The Times Digital Archive* please contact:**

Gale  
High Holborn House  
50 / 51 Bedford Row  
London  
WC1R 4LR  
T: +44 (0) 20 7067 2500  
F: +44 (0) 20 7067 2600  
E: [online@gale.com](mailto:online@gale.com)