

[The Age: national, world, business, entertainment, sport and technology news from Melbourne's leading newspaper.](#)

<http://www.theage.com.au/cgi-bin/common/popupPrintArticle.pl?path=/articles/2009/05/14/1241894111295.html>

16 May 2009

3000 characters in search of content

Prolific contributor Julie Hempenstall (below) has corrected more than 160,000 lines of text for the ANDP.

Photo: *Charlotte Haigh*

May 16, 2009



A remarkable public project is sealing the cracks in our written history, writes Gideon Haigh.

It's understating it to say there was no fanfare when the National Library of Australia rolled out the trial search system for its newspaper digitisation program last July; there was barely the ting on a triangle. Not that the cash-starved library has money for advertising anyway, but the "soft release" was all part of the experiment.

In the international library community, digitisation of old newspapers so that they can be keyword searched is a supercool area. But the NLA's system would be taking this to Brangelina-like coolness by, politely and sotto voce, soliciting members of the public to participate in ironing out the wrinkles in the digital text. What happened next is pretty damn amazing. "What's that movie?" says Cathy Pilgrim, the program's manager. "Field of Dreams? Well, we built it and they did come."

So they - the Australian public - did. Once word began spreading - among genealogists, amateur historians and online library users - the trickle became a flood. Right now, a community of about 3000 far-flung souls are turning on their computers for up to 50 hours a week and tidying text prepared by optical character recognition (OCR) software, as well as subject tagging and even annotating it.

The Australian Newspaper Digitisation Project (ANDP) was always destined to be one of the biggest undertakings of its type anywhere in the world, aiming to make 40million historic newspaper articles available by the end of 2011. But, in the end, it may be distinguished not by its machines but by its men and women.

It is almost a truism that newspapers as we know them will cease to exist in a generation. The irony is that this will coincide with the subversion of an older truism, that today's newspaper is tomorrow's fish-and-chip wrapper.

Old newspapers have always been the most basic raw material of historical researchers. But cycling through pages of microfilmed copies in a darkened room, like the one in the State Library of Victoria newspaper collection, has been for patient souls with steady heads. "I know at least one woman who takes dramamine (the motion sickness drug) before she goes to work," says Pilgrim.

Now technology finally looks like making the fantasy of turning newspapers into searchable texts, intoxicating enough for a host of costly false starts around the world over the past 15 years, into attainable reality - with huge implications for the discipline of history, and the social sciences in general.

Everyone, in fact, wants a piece of it. The European Community, for instance, is throwing money at a project called IMPACT (Improving Access to Text) involving seven libraries, six research institutes and two private sector companies to improve the accuracy of OCR. But at the recent IMPACT conference in Amsterdam, the star was Pilgrim's boss Rose Holley, whose paper about the NLA's trial system bore the brisk practical title *Many Hands Make Light Work*.

To understand how those hands are involved, it's necessary to grasp the basics of OCR.

OCR takes a page of scanned text, breaks it down into columns, paragraphs, words, then characters, gives those characters x and y co-ordinates on the page, then compares them with a set of patterned images on its database. It then gives a confidence rating to its identification of that character, and performs a secondary evaluation of the character in the context of the word using a built-in dictionary.

Even to perform this relatively limited operation, however, OCR must be pampered. The scanned page must have been despecked and deskewed in advance, and the contrast enhanced. And, as anyone who has used early Australian newspapers will tell you, print quality is almost invariably poor, because the earliest presses were English rejects and good paper was in short supply. Because microfilm involves images of bound copies, too, there is almost always distortion down the inside of a page.

When the ANDP began in January 2007, the first stage was a pilot project involving 50,000 pages from newspapers between 1803 and 1954. The result was worrying. "It took three goes to get right," Pilgrim recalls. "It wasn't perfect. Actually, it wasn't even that great." The OCR software was 98 per cent accurate at best, but 71 per cent at worst, the latter implying 145 incorrect characters in an average 500-character paragraph.

Human intervention was essential, for what OCR finds laborious comes naturally to the eye and brain. But who would do it? The NLA has scrimped and saved its way to a \$10million budget for ANDP, but there was no way it could afford correction on the scale required. Holley and Pilgrim credit their lead architect Kent Fitch with the brainwave of involving the public and, although he fittingly calls it a group decision, it clearly represents a long-term aspiration.

Some years earlier, Fitch had worked on the AusLit database, developed jointly by the NLA in partnership with a dozen Australian universities. He tried cajoling management into opening access of the subscriber-only Australian literature gateway so that the public could add content and correct errors, but without success.

"They were very resistant to the idea," Fitch recalls. "I guess because it represented a kind of a loss of their authority, but also because of the prevailing fashion for user-pays solutions, restricting access by charging people for it: you know, the idea that you create something then prove your worth by attracting 'customers'. Although the funny thing was, of course, that this was a system by academics for academics, so most of the money was from the public purse anyway: the money was just going in a circle."

This was the attitude Fitch brought to the ANDP. And this time, with the likes of Wikipedia having blazed the open source trail, he found much greater support.

Unlike a similar digitisation project at the British Library, the NLA committed from the first to a free model. And when the ANDP team members went to discuss beta (pre-release) technology with its core constituency, the denizens of its Canberra newspaper reading room, they found surprising reciprocal enthusiasm.

Fitch has never forgotten that "mind-blowing" visit - the woman, for instance, who was painstakingly documenting changes to the editorial style of The Argus by taking notes on old library cards. "There were all these people here adding value to the content," he says. "But not in such a way that it could be shared."

Fitch particularly recalls a conversation with another researcher who had spent years transcribing the speeches of a New South Wales state politician from the 1930s, building his own huge database.

How would it be, Fitch asked, if that researcher could search for, study and copy online OCR-generated versions of the speeches?

The scholar was suspicious. "What if the OCR was wrong?" he asked.

"Well," said Fitch, "if you could correct it, would you?"

The answer was instantaneous: "Of course I would."

"There are pluses and minuses to open-source technology," Fitch agrees. "But the way we saw it, in every street there's a person who's an expert on something. There's also a huge community of family historians constantly poring over old newspapers."

It was primarily genealogists who answered that first, furtive ANDP call. The beta system allowed them to save corrected text line by line, and also to affix time-saving subject tags. The NLA then added to the sense of cumulative effort by initialising a counter that revealed the most prolific contributors.

Top of the list almost from the start has been 42-year-old Julie Hempenstall from the hamlet of Sutton Grange, near Bendigo, who has spent 20 years tracing her Jewish ancestors back to a London family of 19 children two centuries ago.

Having learned of ANDP on an email list of the Australian Jewish Genealogical Society last July, she searched for the name of her great-great-grandfather. Up came a 1922 Argus death notice. She was hooked. "I'd always planned to go to Melbourne one day to look at the newspapers there as part of my various projects," she says. "Now I didn't have to."

Hempenstall is the beneficiary of another technological advance. Sutton Grange has one of Australia's smallest telephone exchanges, and locals only prodded Telstra into delivering broadband internet in June 2005. As far as the NLA is concerned, Hempenstall has already justified that investment by correcting more than 160,000 lines of text, usually working after driving her daughters to school or putting them to bed.

In the main, Hempenstall enjoys local history, although she also has a fetish for crime. Earlier this year, the NLA sent her a copy of their official history, National Treasures, in appreciation of her efforts. There she read about the infamous Victorian murder Frederick Deeming, the "Jack the Ripper of the South Seas", and promptly identified, corrected and subject tagged 115 articles concerning him.

Hempenstall's sheep farmer partner doesn't really "get it", but she enjoys her regime: she can "look at things that interest me", and also "help other people".

Similar sentiments have been expressed by more than 600 users in responding to a poll and leaving comments on a Facebook page. "What's surprised us is the whole idea of providing a means for the development of social capital," says Pilgrim. "That there's a group of people out there who want to participate in this project for all sorts of personal reasons, but whose main motivation is the benefit of others. One person said it was like applying his watermark to history."

Other respondents gushed: "OCR text correction is great! I think I just found my new hobby!"; "Thank you! You lot are so cool"; "There should be a warning about using this site and its possible addictive effects! I have a great deal of trouble getting back to what I should be doing at times."

One favourite of the ANDP team was Catherine Devine, an Australian based in Washington for the last decade, who corrected because she disliked television but liked typing.

It is too early to determine quite what the ANDP will mean for scholarship, but the possibilities are enormous, even in apparently well-trodden areas, for it significantly narrows the cracks down through which figures of the past have been apt to disappear.

One early adopter is Dr Clare Wright, author of *Beyond the Ladies Lounge* (2003), a history of female publicans, whose more recent area of inquiry is the women of the Eureka Stockade.

Wright has an abiding fascination with Ann Jones, the English widow who ran the Glenrowan Hotel, prosecuted for harbouring the Kelly gang on the night of their last stand.

She was impatient to know what the NLA's system might tell her, and wasn't disappointed.

In an interstate paper that researchers had not previously examined, Wright learned that the warrant for Jones' arrest was issued on the day of Kelly's execution: a hitherto unknown fact, with interesting implications. Was her arrest to be a distraction from the execution? Was it propitiation of Kelly sympathisers who disliked her?

"You would think - wouldn't you? - that every bit of the landscape of the Kelly story had been pored over," Wright says. "But when an interpretative historian gets a new fact, it opens a whole new range of questions."

Wright's happiest moment, however, concerned the fugitive figure of Sarah Hanmer, in whose Ballarat theatre, the Adelphi, the Eureka miners intended meeting after burning their licences - a plan aborted by the storming of their stockade.

Eureka sympathiser Hanmer left Ballarat soon after, disappearing in the process from the usual sources - something Wright is used to.

"In women's history, it is incredibly difficult to follow a thread," she observes. "Women change their names when they get married, and tend not to feature in official records."

But when Wright went looking digitally, who should turn up after Eureka in a Brisbane Courier advertisement for a new theatre?

"Ladies and Gentlemen of the THEATRICAL PROFESSION, Singers, Musicians, Property Man wishing Engagements will please address Mrs HANMER." Subsequent citations verified the Ballarat connection.

There were some misgivings at the NLA last July about effectively rendering up such an expensive project to the ministrations of faceless helpers.

But the ANDP's biggest challenge has turned out not to be flawed, wilful, impulsive human beings, but smart, shiny, expensive machines: the OCR supplier struggled to meet contracted targets, with the result that scanning and OCR work was put out to tender again late last year.

The cost of the project has been unaffected. Indeed, the ANDP is in the decidedly unusual position of effectively getting cheaper as time goes on, processor speeds increasing and cost of storing data decreasing every year.

Costs of the next wave of digitising, to involve the heavily-requested Sydney Morning Herald, have also been defrayed by a \$1 million donation from the Vincent Fairfax Family Foundation.

In due course, that donation will be complemented by an unquantifiable gift from many invisible hands.