COLONIAL READING EXPERIENCE:

A First Report from the Colonial Newspapers and Magazines Project*

Reading Communities and the Circulation of Print conference, ANU April 2014

Paul Eggert

UNSW Canberra

During the last decade research in Australian literature has come to depend on a central bibliographic database, AustLit (http://www.austlit.edu.au). Contributed by a dozen university partners over the years, its extensive contents (about 1 million enhanced records) now have such a broad scope and acceptance among researchers in the field, that it has become the de facto bibliographic authority.[1] Two of the AustLit indexers at UNSW Canberra are here with me today. Tessa Wooldridge and Jane Rankine between them have clocked up nearly 40 years creating and correcting AustLit records. Jane will be driving the machine for me, riffing on the data during the dull bits of my paper rather than always directly illustrating what I am saying.

A book-based focus for AustLit was essential in the earlier years of the collaborative indexing effort if the basic bibliographic mapping of the field were to be completed with the requisite search capacities. A very high percentage of all works of Australian literature published in book form was satisfactorily captured. This focus meant, however, that lesser understood areas, especially literature not published in book form, were accorded selective treatment only.

This especially applies to the literature of the colonial period. Its neglect is the direct and indirect result of the historical operations of the book trade, which had flow-on effects for bibliography as traditionally undertaken throughout the twentieth century – i.e. bibliography considered as being a recording and description of book-based objects and as inherited by AustLit in its early and consolidation years from the 1980s. Although a bibliographic foundation has been laid over the years for the non-book forms of literary expression they are, in terms of their scope and significance, still represented only minimally within AustLit.

---

* I wish to acknowledge the skilled assistance given by Jane Rankine, Tessa Wooldridge and Kent Fitch in the preparation of this paper.

The ungratifying result is simply put. The great bulk of colonial literature has been overlooked by literary historians simply because the principal publishing outlets for colonial authors were the prolific newspapers and magazines of the day. Local book publication was not put onto a proper footing until late in the nineteenth century because the local trade was swamped by the book-export trade emanating mainly from London. Some Australian works were chosen for publication in London, usually because of personal or trade contacts or, late in the century, by agencies of London publishers set up specially in Sydney or Melbourne to secure new titles for the still-successful three-volume novel. Such works made their way back to Australia via this trade route. Although they made up only a minuscule proportion of the whole they are what remained most visible to bibliographers and therefore to literary historians and critics afterwards. Since the late 1980s, work uncovering the phenomenon of serialised novels by Australian authors, very often female, in colonial newspapers extended an earlier listing of poetry in pre-1850 newspapers. Gradually the more general disappearance has been brought back onto the table of scholarly concern.

This has whet the appetite of some literary researchers, first to appreciate the full extent of the colonial achievement and then to interpret it. Methodologically, this is a considerable problem. At the UNSW campus here in Canberra at the Australian Defence Force Academy a significant start on a new project to address the problem has been made. The Colonial Newspapers and Magazines Project, as we call it ('CNMP' for short http://www.austlit.edu.au/austlit/page/5960612), came formally into existence only this year but preparatory work on it actually occupied much of last year. As a result its bibliographic skeleton is already in place, fully articulated. Our effort this year and, if we are given Australian Research Council (ARC) funding over the next couple of years, will flesh the skeleton out. In the past our newspaper data was less systematically gathered, typically being based on the indexing for their Australian literary content of the full- or part-runs of individual titles, their choice usually reflecting the needs of particular research projects or their perceived importance. For instance, with the aid of an ARC infrastructure grant in the late 1990s we completed the indexing of the entire run of the famous Sydney weekly, the *Bulletin*. This gave us valuable data but no representativeness. More is needed for the next generation of research.

Reluctantly, however, I have been forced to conclude that it is not feasible to try to manually index the whole colonial literary field: it is too extensive and the works too numerous, making the indexing too

expensive to achieve comprehensively. Thus, systematically targeted sampling is the only practical alternative. A chronological method has been arrived at after much discussion and preliminary investigation.

Consequently, a listing of all those newspapers and magazines that were being published in the Australian colonies in three particular years, 1838, 1868 and 1888, has been definitively established. This process, which we naively believed would be straightforward, proved to be anything but, and took many months of Tessa's time last year. The increase in the number of titles over the period is surprising: predictably modest at first (36 titles in 1838), only 50 years after first settlement, 218 by post-goldrushes 1868, and continuing to rise strongly again to 524 by the centenary year, 1888. 'Parent' records for all of these newspaper and magazine titles was completed in 2013, funded by ARC and university contributions. (Parent records give the history, editorship and high-level publication details. Then, under the parent record, individual issue records are collected. They in turn 'contain' records of all of the relevant items published within each issue of the newspaper.) This database organisation will permit a full specification of the literary contribution in these years.

Because of some generous UNSW Canberra funding for 2014, by the end of the year 63 of the 778 titles will have been indexed to CNMP standards. With the anticipated ARC funding in 2015, for which we have just applied, and a promised UNSW contribution to go with it, another 90–100 will be completed. If funding becomes available also for 2016, a similar sizable proportion of the remainder will then be indexed. The coverage, for the three selected years, will even then not be complete. Calculations were perfomed, based on how long the indexing had taken on average in our preparatory phase last year and early this year. They led to the unwelcome conclusion that only a third of the total titles for the selected three years can be covered. Nevertheless, this authoritative sample will cover the main dailies in the colonial capitals, the bulk of the monthlies, quarterlies and annuals, and the weeklies and bi-weeklies of the principal country towns. (This last is important, since the majority of the population lived in the Bush and the literary content was often considerable in the weeklies, which were the most common format). The diversity if not the full extent of the colonial literary field will thus be captured. The year 1900 will, once again given funding, be similarly approached in an envisaged third year of this project in 2017.

These large samples, indexed to an authoritative standard, will thereby give us a firm base for reliable statistical and other extrapolations in all of the fields that AustLit's records populate for each work and each

agent (by 'agent' we mean author or organisation). Some envisaged technical developments that I will describe in a minute will, if successful, ultimately permit a semi-automated means of gathering the literary data. The level of reliability of that data will be able to be tested against, and corrected for, the authoritative data from the traditional indexing, thus enabling broader characterisations of the whole colonial period to be contemplated.

So far I have probably given you the impression that, as with the rest of the AustLit database, the data capture is, as in the past, for works of Australian literature only. Now it is true that we have become more flexible and generous in terms of those genres that may be considered 'literary', and we have included works by short-term visitors to Australia provided they have Australian subject matter. Significantly we have always included reviews and essays *about* works. Subject terms are also provided for each work indexed. Thus, searching on subject terms, year-ranges, genre, place, publisher and on works about other works can easily be done, with each search parameter able to narrow the other search parameters if desired. But what AustLit has not included until now is the reception of non-Australian works. If we were to try to get a grip on the colonial literary and theatrical experience it seemed to me that AustLit's traditional delimiter had to be relaxed. We had previously treated works by, say Shakespeare, Luther and Goethe as being outside AustLit's remit, even if performed, sold, reviewed or discussed more formally within Australia: that is to say, if given a presence or life within a particular cultural setting.

You see which way I am going here. Since the early 1990s book history and print culture have been gradually sensitising us to the phenomenon of reception as being a potent dimension of literary works. From my viewpoint as an editorial theorist, such receptions need to be incorporated within the definition of the work, not just theoretically but also empirically. But the question was: Should this emerging awareness presume to dictate the multifarious workings of a large database that has been so expensively produced over 25 years now? I came to the conclusion it should, and AustLit's manager Kerry Kilner agreed. Although foreign works and authors are not accorded the same full treatment as Australian works receive they are now being indexed for the CNMP nevertheless and are thus open to searching. Most users of AustLit, perhaps unaware of the CNMP, will normally be searching for specifically Australian data. For that reason, foreign works and agents are given the tag 'international' and can be excluded from ordinary searches, and also vice versa.[2]

An inevitable result is that periodicals that were indexed before the CNMP project to a tighter definition of what Australian literature encompasses – and there is quite a list of them – http://www.austlit.edu.au/austlit/page/5962135 – won't provide all the data we need. One obvious question to be asked is the extent to which the colonial literary mentality was conditioned by British and other countries' literatures, and by their plays and songs, as opposed to its conditioning by *local* newsprint and stage productions. The data collected in the past won't get us very far with answering this question. This simply reflects the fact of life that research agendas don't stand still. In the very early years of the old AUSTLIT in the 1980s what we later came to call 'life-writing' was deemed to be outside or on the margins of literature as defined for database purposes. If insufficient funding meant any literary form had to be left aside for later it would be life-writing. Travel writing was also considered secondary. Such absences are regrettable and, as I have said, later the boundaries were relaxed.

So it is that with the CNMP we are widening the indexing mandate to acknowledge the new kinds of literary- and print-historical questions that 'big data', as it is nowadays called, may help us to conceptualise and then begin to answer, not via *book*-based sample and anecdote as predominantly in the past, but through data that is more representative of the field under investigation as a whole.

So far I have told you we aim to do this through three or four nicely spaced years across the 113 years that made up the colonial period. Then I added that even with generous funding only one-third of the known newspapers and magazines of those three years will likely be indexed. So how might we bridge the yawning gaps that these figures highlight?

Semi-automatically identifying literary content may well be feasible. You probably know about the National Library's ongoing program to digitise all Australian newspapers. The database, called TROVE, has processed about 12 million pages so far. Typically, for TROVE, an existing microfilm of a newspaper is digitised to provide a facsimile of the pages, and the images are then OCR'd, producing text of varying legibility.

Digital humanists tell us that we are 'cognitively biased towards the unique'.[3] If we are studying a historical document or poem or film we expect it, in the form we encounter it, to be a reliable focus for our interpretative activity. However, if we are dealing with very large numbers of such objects and we are trying to identify and then interpret overall trends a certain amount of unreliable data is tolerable.

But the question is, how much? My extrapolations from existing data suggest that there may have been more than a quarter of a million literary works produced in the Australian colonies during the colonial period. If that's true then we can probably tolerate a 5% error rate but not 95%. We can tolerate 5% because that number will not necessarily change the patterns that the statistics otherwise reveal. So our data only needs to be *reasonably* reliable for the purpose in hand, even though *highly* reliable would make us feel more confident.

Although it has been ascertained that about 18%–20% of words in ocr'd text of 19th-century newspapers have one or more errors in them, recent algorithmic experiments, whose results are to be announced in Madrid at the DaTECH conference next month, show an error rate of around 7% is achievable. Kent Fitch, the longstanding programmer for AustLit and John Evershed are the authors of the paper and devisers of a complex algorithmic system called *OverProof* claimed to be able to achieve this rate.[4]

Thus it should become possible, using analytical techniques that search for literary markers including 'fuzzy' matches on names and titles, to generate collections of newspaper items that can subsequently be examined to determine their relevance for AustLit – in other words, to produce a semi-automated indexing. The level of reliability of the data will be able to be tested against, and corrected for, the authoritative data derived from our traditional manual indexing. The data will *also* allow us to begin to arrive at statistical answers to more general questions about the colonial literary experience: its productions in newspaper form, say, in comparison to its book forms, locally written works as opposed to imported ones, Melbourne productions vs Sydney, city versus the bush, the relative presence of American literature and plays, and how any or all of these phenomena changed over the colonial period – as you anticipate they would have done, that is, if you still accept the nationalist case about the 1890s and the role of the *Bulletin* in it.

The generation of such questions and then the evaluation of the returns will be the role of the interpreter.[5] Clearly such work falls within the recent international shift in literary studies towards distant reading via collections of data of various kinds, and in *that* it is good to be able to play a role.

Of course I realise that to be able to *show* that a mapping of the colonial cultural mentality is feasible will take more than what I have argued today. It will take experiment, failures no doubt, followed by new designs and new questions. Better-defined boundaries around the ambition and about the competence of the data to serve as meaningful

evidence of larger claims will clarify in time. Nevertheless, if it can be pulled off it would be a major research achievement and would open up all manner of enquiries, of importance for the areas of literature and print culture internationally and perhaps to other areas as well. As far as I know, no other country has a literary database with the breadth and depth of bibliographic indexing, and the sophistication, to permit such a project to be built upon it.

Surprises in the manual indexing so far

Now I need to give you a taste of the surprises that the manual indexing has been throwing up recently.[6]

Let's take Sydney in 1838. Its white population, including the rest of the settled areas fanning out from Sydney, was only about 90,000 or so.[7] By that year the most important of the colonial-born poets Charles Harpur had been publishing in the press for five years. Jane's indexing of the *Sydney Gazette* for the first six months of 1838 reveals that the majority of items of a broadly literary or print-cultural nature had to do with the stage. From 1 January to 12 June 1838, there were 70 advertisements for plays, 15 advertisements for book sales from booksellers and auctioneers and 10 advertisements for printers and journalists. There were 21 reviews of which 17 were for theatrical performances. Of the columns indexed 19 are on the theatre, 11 are about the newspapers of the time including threats of libel etc., 9 are about book imports by ship, and the rest about cultural subjects such as the 50th anniversary of first settlement. So far Jane has found only 2 poems, 14 short stories, 3 extracts and a serialisation of the *Pickwick Papers*, an appropriation still believed to be legal in 1838 but not for much longer.

I'm only skimming the surface: if you want to know what the Swan River colonists, all 2,000 or so of them, were reading in 1838 or if you want to know which Shakespeare plays were available to Sydney colonists that same year, or which Voltaire volumes they reading, you can search for yourself. The interface is gradually being improved to facilitate print-culture as opposed to more purely Australian-literature-type enquiries. If you get into strife while searching you can always email Jane or Tessa for help, as we are all interested in what users will make of this extension of AustLit's bibliographical service.[8]

One of our past indexing activities at ADFA, carried out from 2006 to 2009 by Kay Walsh when some new funding unexpectedly came our way, was to write some hundreds of entries for publishers and printers who were known to have been active in some way in the Australian literary scene, particularly in the colonial era. We were trying to enrich

the existing agent records with information that would potentially put the works produced by these publishers and printers in a new light and allow networks of agents and works to be plotted. You'll appreciate that AustLit is not a bare-bones bibliographical source. Some booksellers were covered by this project and very often an outline biography was appended. Many other booksellers will get the same treatment as we accumulate information about their advertising: this is part of the CNMP project and will become more powerful as we proceed. The decision to index booksellers' advertisements – something we would not have even considered in the early days of AUSTLIT – should finally allow telling patterns to emerge.

Tessa has been indexing Henry Parkes's daily, the *Empire* for 1868. By now, the NSW population was nearing half a million. Predictably, monthly summaries of the English periodicals were being provided in the *Empire*, as were (just as in 1838) announcements of the arrival of books by steamer, now including via Panama even though the canal would not open until 1914. Predictable also were advertisements for the Sydney Mechanics School of Arts (established in 1833 and soon with its own library) and from the Australian Library (from 1827). The latter, we learn, had 600 subscribers and over 20,000 volumes by 1868. The Burwood Literary Institute also received a report on its activities in the *Empire* and the Working Men's Book Society advertised volumes for sale. Less predictable perhaps, are a small number of reports of lectures on literary topics and a profusion of reports and reviews of recitals by the Scottish elocutionist Miss M. E. Aitken, which were evidently popular. She gave many recitals, including one on 25 April 1868 'in compliance with a Special Request conveyed by a Deputation from the Working Classes'. An advertisement gives full details of the program (http://www.austlit.edu.au/austlit/page/7080312).

A keyword search on the CNMP for the term 'literary readings' at the moment produces 82 results for the period January–April 1868 (http://www.austlit.edu.au/austlit/search/page?facetSampleSize=10000&facetValuesSize=10&passThru=y&count=50&agentQuery=&workQuery=((literary readings)) AND ((wdate%3A1868)) AND ((waffiliation%3A"CNMP"))&pseudoscpe=work - keywordlliterary readh). So-called Penny Readings, whose pricing implies a working class audience, were a phenomenon of the time. The Surry Hills Young Men's Mutual Improvement Society and the Union Literary Club seem also to have catered to this taste. Coincidentally, this was the year of Charles Dickens's tour of America, with readings from his own works, a tour that was duly reported in the *Empire* from the *Atlantic Monthly*. Samuel Taylor Coleridge's grandson Derwent Moultrie Coleridge performed a reading of Dickens's *A Christmas Carol* in Woollahra in Sydney during 1868 (http://www.austlit.edu.au/austlit/page/6943827). In this period just

before the introduction of compulsory education it may be that literature was still being more listened to than read. At the moment this conclusion is guesswork. Statistics may finally reveal the trend, once we have enough data.

Judging by reviews and reports in the *Empire*, there were 55 different plays mounted on the Sydney stage during the early months of 1868. Of the 55 only three were definitely Australian. Shakespeare was the third most popular in terms of the number of plays. He was pipped by the sensation-dramatist Dion Boucicault, whose follower, the Australian Walter Cooper also had one play mounted called *Colonial Experience*, and by Tom Taylor who had seven plays mounted. Taylor was the author of *Our American Cousin* (1852), the play that was being performed in the presence of Abraham Lincoln when he was assassinated in 1865 (http://www.austlit.edu.au/austlit/page/7232374). Now we know it was performed in Sydney in 1868, along with another six of his plays that same year. The American notoriety cannot have done him any harm in the Australian colonies.

Conclusion

The old AUSTLIT was launched by the former prime minister Gough Whitlam in the bicentenary year 1988 in the ADFA Library. The new AustLit, a consortium of university partners that absorbed the old AUSTLIT, began in 2001. Last year we celebrated 25 continuous years of AustLit indexing at ADFA when, more or less simultaneously with the celebration, the present plan for the CNMP was taking shape. I wonder what the next 25 years will bring for the study of Australian literature, and literature and print culture, in Australia and internationally?

---

[1] Google Analytics show that, during 2013, AustLit received 438,000 visits from 319,000 unique visitors who used 1.9 million page views for an average of 4.4 pages per visit, spending on average 3 minutes 22 seconds per visit.

[2] We also had to come up with a new indexing template to encompass the form 'advertisement' for this project, something you don't do lightly in a database as big as AustLit given the knock-on effects for the rest of the data and for other projects that deposit their data within AustLit.

[3] Jonathan Hope and Michael Witmore, 'The Very Large Textual Object: A Prosthetic Reading of Shakespeare', *Early Modern Literary Studies*, 9.3 / Special Issue 12 (January 2004), 6.1–36 <URL: http://purl.oclc.org/emls/09-3/hopewhit.htm>

[4] For *OverProof* see http://overproof.projectcomputing.com

In computer science there is a body of research going back to 1964 into how ocr'd text can be rendered more accurate. It has been ascertained that the error rate for newspaper ocr'd text, especially that dating from the nineteenth century is around 20%. That is, 20% of words have one or more errors in them. Most ocr'd errors are egregious ones and readily identifiable as such. To the human eye such error is not usually a big obstacle because we can usually pick the word that was intended. But the computer, alas, has a stubbornly pedantic streak and needs to be explicitly taught language modelling in order to make the leap that human readers make intuitively. The computer's advantage is that it works extremely fast with probabilities. For instance that *this* ocr'd set of characters ought to be *that* set of characters given that the latter appears elsewhere in significant numbers and usually in combination with other predictable words. These are called n-grams: e.g. 'the Prime Minister of Australia' is a 5-gram. If the ocr'd text identifies all but one of the words correctly there is a high chance that the fifth word may be correctly identified from the n-gram. This is but one form of probability.

Algorithmically, probability may be built upon probability, the one making the other more powerfully predictive. What is called error modelling can also help if, by appeal to what computer scientists call a 'ground-truth' language corpus, (in effect, texts that have been checked carefully), the spelling *Australla* may be deemed to be probably *Australia*. The town name *Kalgoorlie*, for instance, has over 100 ocr variant spellings in TROVE, and one of the seven most common variants of the name returns over 45,000 hits. None of these will be returned by a search on the conventional spelling. Ingenious algorithms, in combination with one another, lower the probability of error by identifying the unlikely reading and then substituting the more likely one. The very latest research is to be presented at the DATech conference in Madrid in May 2014 by John Evershed and Kent Fitch. Kent also works as the programmer and software engineer for AustLit. After six years of working at the problem, the two of them have now designed an unsupervised ocr correction system called overProof

(http://overproof.projectcomputing.com). In their Madrid paper they demonstrate it is able to reduce the 20% error rate by over 65% and what they call search-misses and false positive returns by 60%.

[5] More local questions may be able to be answered too: such as, Did the 1842 Copyright Act of the British parliament, as has been thought, have the effect of preventing the importing into the colonies of cheap and foreign editions of British works so as to safeguard the copyright of the British author? Comparing data of 1838 with 1868 may yield an answer.

[6] For instance, what was the situation in Perth in the year 1838? In 1836 when a census was taken, seven years after first settlement, a grand total of 1,958 white Europeans inhabited the Swan River Colony that would become Western Australia, a land mass of a million square miles, ten times the size of the British Isles. By 1843 the white population had grown to 3,853. So in 1838, nine years after first settlement, there would have been perhaps 2,500 people. What were they reading? And how did it get there, and how circulated? Were there theatres yet? Jane's indexing of the *Perth Gazette and Western Australian Journal,* which was established in 1833, has revealed its publication of 12 creative works in 1838: 5 poems, one of which is, in the term we use for non-Australian works, 'international'. Entitled 'The Sovereigns of England' its authorship is unknown but it was anthologised in various international periodicals. 2 items are prose extracts, one of which is a piece by Sir Walter Scott, 'Intelligence in Card Playing', which may reveal a certain taste among the colonists. There is one short story. More revealing are 4 notices for the Western Australian Book Society, which had been established in 1835. On 20 January 1838 and 11 August 1838 society members are advised that books have arrived and are ready for distribution. Presumably the society was importing them specifically for members. Rivalries amongst newspaper proprietors are a feature of the colonial press. It started early in Perth, with the Gazette, regularly attacking the opposition newspaper the Swan River Guardian, whose proprietor William Nairne Clark was pushing for an independent press unaligned to the government as the early gazettes generally were.

[7] Censuses gave figures for the white population of 77,000 in 1836 and 119,000 in 1841.

[8] What of 1868? Tessa has been indexing a daily newspaper that was one of most prolifically literary ones, Henry Parkes's the *Empire*. This sonorously loyal name actually disguises how attentive it was to local productions. Tessa's report to me on 4 February this year, when she had just finished indexing the *Empire* up until 1 February 1868, is worth quoting. She is describing a single issue for 1 February 1868 (http://www.austlit.edu.au/austlit/page/C807518):

> The issue contains 15 items indexed for AustLit: 7 advertisements, 4 columns, 1 short story and 3 poems. (One poem only had been indexed previously.) This issue took 2 ½ hours to index. The single advertisement for Maddock's Select Library: http://www.austlit.edu.au/austlit/page/6973000 took one hour. Why? The creator of the advertisement was not on AustLit; I needed to research the creator and create an agent record; in the course of that research, I discovered an important travel book, published in 1888 that was not on AustLit (its author discussed Maddock's Library in the book [viewed online]); the book's author was not on AustLit; I created records for the book and the author; finally, I set up the record for the advertisement. Simply indexing one advertisement is very quick, but the background work required can take a long time.