Fiction in 19th-century Australian Newspapers on Trove

Australian Historical Association panel
Wednesday 4 July 2018; 11:00am-12:30pm
The project I want to talk about today seeks to discover and explore fiction in 19th-century Australian newspapers. These newspapers are of literary and historical interest because they were the main source of fiction for readers in this time and place, and the main avenue of publication for local authors; and hence, the main vehicles in the development of Australian literary culture.
This project is possible because Trove offers an excellent – and the world’s largest – open-access, mass-digitised collection of historical newspapers. This digitised collection makes it possible to explore the fiction published in Australian newspapers for the first time in an extensive and systematic way.
I’m planning to discuss three issues relating to this fortunate conjunction of literary historical circumstances and contemporary remediation:

- How I discovered fiction in these newspapers – focusing on what I see as some general principles for large-scale historical research;
- Some of the contributions to literary and cultural history that these findings offer;
- Ongoing work to use large-scale data as the basis for scholarship as well as new collaborations and publics for historical research.
Here’s a diagram of the process by which I discovered and described this fiction.

It includes some fairly standard features for research at scale.
It uses Trove’s API or Application Programming Interface; which anyone can use to export data in machine-readable form.
The API stores the results in a database with an editable interface (and I’ll demonstrate and discuss that interface in the final part of my talk).

I want to spend a bit more time with the two elements of this sequence that are particular to my project at the same time, as they have broader implications for historical research at scale.
The first of those elements is Step 1: Paratextual keyword query.

We know that we can search Trove newspapers using keywords to find articles of interest; by that approach, to discover fiction it might seem logical to search for authors or titles of fiction. But this approach has a number of significant problems:

- Have to progress author by author or title by title;
- You’d have to exclude a lot of irrelevant results – obituaries, book reviews, literary gossip;
- More than simply being time-consuming, this approach would reinforce prior conceptions of the authors and titles published in Australian newspapers in that you’d find the ones you expected to be present;
- And it wouldn’t accommodate the widespread anonymous publication of fiction or changes in titles as stories were published and republished.

The solution I developed I call the paratextual method. Takes advantage of the very high consistency of paratextual features in newspapers – the words used to introduce and describe articles. On this basis, it uses words such as “serial”, “story”, “novelist”, “our author”, and most usefully “chapter” to discover stories in the newspapers. The approach works because Trove zones pages into articles and defines and corrects title information, meaning that the paratextual information is both identified and more reliable than the rest of the OCR rendered text.
Because of the way the relevance ranking algorithm works to privilege articles where words appear in the title and are repeated throughout this approach meant that the vast majority of the initial thousands of results for searches with these different terms were fiction.

**Take home for historical research at scale:** methodologies need to be devised that combine disciplinary knowledge – in this case, of the generic and documentary context being investigated – with knowledge of the technical protocols, features and affordances of the digitised collection; either one without the other is insufficient.
The second element I want to highlight is Steps 5 and 8: Automatic, semi-automatic and manual exclusion and correcting of records.

Interaction of this paratextual method with features of Trove’s digital collection made it possible to export fiction. But the harvested form is not especially useful for literary historical research.
This is an example of a record and you can see, for instance, that there’s lots of different types of information in the “<heading>” field; but it’s not structured in a way that is generalisable to other articles.

ARC enabled me to hire a bibliographer, Carol Hetherington, for 3 years to work on extensive manual processing of the harvested data to render it as useful literary historical – bibliographical – information.
This involved, for instance, extending the information in Trove’s “heading” field into multiple separate categories including title, author, copyright and circulation information, chapter number and chapter title; and then descriptions of those categories, such as the form of authorial attribution.
It also involved incorporating additional information about fields, for instance, whether a newspaper was metropolitan, suburban or provincial; consistent location information; and stitching multiple different banner changes together under common newspaper titles.

Finally, there was extensive other information gleaned through bibliographical research (such as the common title or author name uniting the different titles and attributions in the newspapers; and details about gender and nationality of authors; sources of information and other sites of publication).
And after all that there was still an extensive process of editing and correcting the bibliographical data.

**Take home for historical research at scale:** creating data that is useful for humanities research is not a simple matter of sucking information out of Trove and analysing it for historical trends; prior to such analysis, significant research is often required to produce useful data; and this adds enormous value to what Trove currently makes accessible.
As a result of a harvest conducted in July 2015, when I estimate that around 30% of 19th century newspapers had been digitised, we discovered over 21,000 novels, novellas and short stories. We found that about 80% of the 19th century newspapers contained fiction – large amounts in most cases.

The numbers are not of the scale associated with so-called big-data research, but when you consider that, before this project, the bibliographical record for what we knew was published in 19th century Australian newspapers was in the hundreds this is a significant increase on existing knowledge.
Once we conducted analysis, we made discoveries that significantly transformed prior understandings of Australian literary and cultural history:

- Rather than these newspapers focusing largely or entirely on British fiction, as has been argued, this was a cosmopolitan literary and reading culture;
- There was much more Australian fiction than previous histories had proposed, with the discovered titles including hundreds of new titles by known Australian authors as well as hundreds of titles by over 60 previously unknown local authors (with the likelihood being that lots of the fiction where we’ve been unable to determine the authors’ identity was Australian);
- There’s evidence to suggest that this Australian writing was accorded significantly more value than has been argued, with the paratexts consistently highlighting local authorship, and there are some examples of how that occurs.
SOME DISCOVERIES ABOUT FICTION IN COLONIAL AUSTRALIAN NEWSPAPERS

• Consistent, prominent depiction of Aboriginal characters in local fiction – the “bush tradition” (but not as we’ve known it)
• New structure and organization for 19th-century Australian literary culture – cross-colonial provincial newspaper syndicates were the major publishers of Australian fiction and fiction in Australia

The local fiction also challenges a widespread argument in Australian postcolonial studies that 19th-century fiction replicated the legal lie of terra nullius with Aboriginal people absent from Australian stories. In fact a characteristic feature of the Australian fiction was its consistent and prominently depiction of Aboriginal characters. The word cloud on the slide signifies the topic deemed by a machine learning method to be most strongly predictive of Australian writing – if you look closely you can see words indicative of the presence of Aboriginal characters (black, blacks, track, mob) alongside more rural settings, characters, and activities

The bush tradition lives – but not as we thought we knew it. Perhaps most remarkably, we discovered a new structure and organization for 19th-century literary culture: “bush tradition”. Most previous histories have focused on books and metropolitan newspapers. In fact, provincial newspaper published more than either of their metropolitan counterparts, and did so in highly organised ways.

Exploring incidences of reprinting in the dataset shows that much of the provincial fiction was supplied by an extensive array of fiction syndication agencies, both Australian and international, that operated across colonial lines such that readers in outback New South Wales were reading the same stories as their Victorian, South Australian, Queensland, Tasmanian and sometimes their West Australian and New Zealand counterparts.

This cross-colonial provincial culture challenges the view, in Alan Atkinson’s words, that the Australian colonies operated as separate “nations” which “each of the Australasian capital cities ... the centre of its own world,” including its own print culture. Many of these syndicates focused on
Australian fiction and they emerge as the major publishers of such writing certainly in the 19th century and probably up until at least the 1970s.

The prevalence of local fiction in provincial newspapers from the late 1870s also dismantles a key tenet of Australian literary history, which locates the origins of local literary culture in the 1890s in the “Red Pages” of the Bulletin magazine; in fact, this Sydney-based publication followed in the footsteps of the provincial syndicates.

Let me move, now, beyond my findings for the project to consider how large-scale projects can enable other scholarship as well as new collaborations and publics for historical research. Large-scale historical data takes such a long time (and often so much public money) to create and embeds so much important knowledge that I think it’s essential we publish our data wherever possible.

The way in which we publish has implications for how it is used; and over time I’ve moved from what I now recognise as a broadcasting model of publishing to a collaborating one that not only makes data available but improves and extends it while facilitating sustainability of research findings.
All 21,000+ novels, novellas and short stories that we discovered are available in an editable database connected to Trove via the API (access it using that URL).

The database allows me to publish the data underpinning my arguments – so that others can explore, contest, confirm, or extend them.

In data-rich humanities research, data publication tends to involve putting that data in an open-access repository, which has the benefit of the researcher not having to act as publisher, but the drawback of the data being effectively available only to those who have the computational skills to wrangle large datasets.
In a database, it offers a resource for research, whether it is computational or non-computational; there are various facets to limit searches as well as keyword searching, and if you’ve got good eyes you might be able to see that I’ve limited the search here to Canadian women writers – and we can see that romance writer May Agnes Fleming is very well represented in that category.
You can explore the individual instalments or chapters of the fiction. You can also export textual and bibliographical data for computational analysis, whether of the whole database or a defined subset.

These stories could certainly provide a resource for traditional research and general publications.
There’s my book, on the left, based on this collection and on the right is a collection I edited with Canberra’s own Obiter Press last year publishing five newly discovered stories by important 19th century Australian feminist, socialist and novelist Catherine Martin. This year we’re working, with different editors, on two further collections of 19th century Australian Xmas stories and mining fiction.

However, I am also exploring how data publication can enable new forms of collaboration beyond the academic and publishing spheres. Because the database is connected to Trove’s API, members of the public can find a story they’re interested in in the database...
... correct the text in Trove and ...
... draw that corrected text back into the database, thus improving the accuracy of keyword searching in both sites.
They can also add missing chapters or instalments that they discover in the digitized newspapers, or entirely new titles as well as to...
...publish the completed and corrected stories.

Since the database was launched in March this year I’ve had over 400 members of the public sign up to the research community: many focus on text correction but a substantial number add new stories to the database, often by focusing on particular authors or previously undigitized newspapers. Last time I checked over 300 new titles had been added and many more than that had been corrected.
A further collaboration is with Trove itself. I worked with the Trove team to link our two databases so that they could harvest all of these records of fiction and represent them in Trove.

Now, when someone searches the National Library of Australia’s catalogue ...
They'll find records of fiction in Australian newspapers including stories where Trove indexes other versions: here book editions...
As well as stories where the only record in Trove is from the digitised newspapers.
Clicking on harvested records shows the detail, including links to the digitized newspaper articles that comprise the instalments of the stories and to the Australian Newspaper Fiction Database. This connection with Trove completes what we might call the humanities data cycle, in that the project both draws on and feeds back into a national digital collection, thus enriching knowledge and accessibility of the documentary record, while also ensuring the sustainability of research findings.

In collaboration with some colleagues in education I’m also hoping to develop a project wherein this data publication can act as a resource for teaching in subject English in Australian secondary schools that seeks:

- to bring the history of reading (of Australian literature and literature in Australia) into conversation with new media literacies (such as the nature of OCR, how collections are created); and
- enables students to act as independent researchers and curators, while engaging in the discussion of literary works that has typically comprised the treatment of literature.

I hope what I’ve shown today are some of the ways in which large-scale research in the humanities is draws on existing practices and knowledge – relating to the historical context we’re exploring, the types of arguments relevant to our fields, and existing collaborations of academics and publishers – but it also enables new practices and knowledge – including new collaborations – that have the potential to revise our understandings and connect us to new partners and publics.