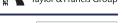Routledge
Taylor & Francis Group

ARTICLE

Check for updates

Metadata as a machine for feeling in Germaine Greer's archive

Millicent Weber[a] and Rachel Buchanan[b]

[a]School of Literature, Languages, and Linguistics, Australian National University, Canberra, Australia; [b]Independent scholar, Melbourne, Australia

ABSTRACT

What happens when a human coder meets a machine one? This article explores this question with reference to the archive of Professor Germaine Greer: Australian-born feminist, performer, scholar, and professional controversialist. It does so by staging two very different data encounters with the 70,000-word finding aid for the print journalism series, a key component of Greer's archive. The first encounter is archivist's creation of the finding aid; the second, archivist and literary scholar's interpretation of this archival metadata using sentiment analysis. Interrogating these activities side-by-side opens up a productive middle ground between humanities scholars and computer technicians, between historians and archivists, between the hand made and the machine made.

This article argues that sentiment analysis offers a new and highly productive method of interrogating archival metadata, and that, as a method which privileges emotive understandings of content, it is particularly appropriate to the study of feminist archives like Greer's. It also argues that these kinds of detailed finding aids are new datasets that reward analysis in their own right, and particularly when considered in dialogue with—rather than simply used as straightforward navigational tools for—the 'original' archival content.

Introduction

Germaine Greer is the author of *The Female Eunuch* (1970),[1] one of the most widely read texts of second-wave feminism. She continues to enjoy a high public profile in Britain, where she has lived since 1964, and in Australia, where she owns a rainforest. The Germaine Greer Archive, purchased by the University of Melbourne in 2013, is exceptional for several reasons. Most collections of personal papers in Australia occupy about six boxes and most of them document a moment or an activity – war service, writing a PhD, setting up a farm – in the life of a man. The Greer Archive currently comprises 500-plus boxes, and it documents all aspects of the personal and professional life, from age 18 onwards, of an Australian-born woman who is a feminist, scholar, journalist, performer and environmentalist.

CONTACT Millicent Weber ✉ millicent.weber@anu.edu.au 🏫 School of Literature, Languages, and Linguistics, Australian National University, Canberra, Australia

The Greer Archive is exceptional for less obvious reasons too. The 33 series that were made available to researchers in April 2018 are encased in more than a million words of rich, carefully considered, descriptive metadata: a map for a metropolis. The University's desire to manage risk and improve discoverability of Greer's records, and the financial support of a small group of donors, lead to a shift back to older ways of working: the archive has been described, in detail, at item level; a rare practice in an era where minimal-level processing is the norm in archives, libraries and museums around the world. These finding aids enable keyword searching. They also open up possibilities for other forms of interrogation. They are digital, richly textual, and extensive, together totalling close to a million words. And these features mean they can be read using computational techniques.

This unprecedented collection of metadata is the subject of this essay. The co-authors use personal reflection and computational techniques, specifically sentiment analysis, to interrogate one corner of the Greer metadata metropolis: the 186-page list that describes 1268 paper items (files) that make up the Print Journalism series (2014.0044). These approaches, one computerised and quantitative, and one human and qualitative, are put in dialogue, sometimes oppositional and sometimes co-consti-tutive. The practice-led, reflective component of this research supports critical analysis of the decisions that went into writing the print journalism finding aid: the tug between objectivity and subjectivity, the pull of Buchanan's personal history in print journalism, the problem of historical language and subject headings, the hate, horror and humour of individual records. This is juxtaposed with analysis of the findings from the senti-ment analysis process. Together these methods reveal how the metadata contains multiple types of evidence not only of Greer's career as a print journalist but also of the career and life history of the archivist who created the metadata and the ambient knowledge developed within the team she led. These methods also resonate with the affective turn within fourth-wave feminism[2] that has started to inform the politics of contemporary archival practice.[3]

Background

In archives, libraries and museums, what Eira Tansey describes as a 'multi-decade cycle of poverty' means metadata is much thinner than it used to be.[4] Last century, archivists laboured to create granular levels of arrangement and description at item level. The detail contained in this information was useful to researchers, but generating these descriptions was slow work. The backlog of hidden, uncatalogued collections grew. This century, archival processing has 'dramatically shifted towards minimal-level processing'.[5] This has become so ubiquitous that it has its own acronym: MPLP, short for 'More Product Less Process'.[6] MPLP collections are generally described in summary at the collection, series, box or file level rather than in detail at file or item level and are minimally indexed, if at all.

The MPLP approach means collections of personal papers at the State Library of Victoria are accessioned and made available to researchers with only the most skeletal of box lists. The handful of records at the National Library of Australia that have been fully listed are small – for example, the papers of author Patrick White occupy just over six metres of shelf space;[7] those of librarian Sir Harold White about 22.[8] As Graham

Powell noted over 20 years ago, about 90% of personal collections in Australia comprised no more than six boxes – about 1.5 metres of shelf space.[9]

The Greer Archive sprawls across 82 metres, and the entire archive has detailed file-level metadata supported by subject indexing. If metadata is a map, then there are few, if any, unmarked streets in the Greer Archive.

A desire to manage risk and improve discoverability of Greer Archive records – and the financial support of a small group of donors – led to a shift back to older ways of working, including the creation of metadata in detail at file level rather than summary file or box level.[10] A further innovation is the archivists' construction and use of 'the Greer Thesaurus', a working document that attempted to create a standard language to describe material across the 33 series, using a mixture of Library of Congress subject headings (LCSH) and non-LSCH terms to capture key terms (2018.0054.00008). The use of standard language for subject indexing is common in libraries but appears to be less common in Australian archives, although many national and local standards recommend the use of thesauruses to select and formulate subject access points or index terms. This perhaps reflects the difference between published and unpublished material including the greater variety and unpredictability of archival content within single units of description as well as different traditions of archival description.[11]

But what is all this writing for? Are the Greer lists merely very detailed navigation charts for a collection of records, 'an aid to discovery and retrieval', as Chris Hurley puts it?[12] Are they democratic descriptive surrogates – discoverable online via Trove, the UMA database or a basic Google search – that stand in for a collection of paper records that is Restricted Access and can only be read in a supervised reading room? Or are they also records of recordkeeping activity of the archive's creator, Germaine Greer, and of the archive's new custodian, the University of Melbourne Archives? If the Greer Archive metadata is a record of recordkeeping activity – and the series scope and content notes certainly are – then who is the primary audience for the metadata? Researchers? Germaine Greer? Other archivists?

This article considers these questions by using a computational technique called sentiment analysis to interrogate the metadata Buchanan created for the Print Journalism series, a collection of paper records that documents Greer's international career as a journalist (1959–2010). Listing this series was a demanding task that required a blend of mundane data entry and simultaneous content analysis of pieces of writing that are complex, diverse and often confrontational.

Buchanan, the curator of the Greer Archive between October 2015 and March 2018, listed eight other series in the Greer Archive, including Ephemeral Publications (2017.0010), Internet (2018.0055), Contraceptives, Cars and Gardens (2017.0026) and the Early Years Academic, Performance, Writing and Personal Papers (2014.004) series. Although both the Early Years and Print Journalism series are relatively small in physical terms – 20 boxes and 24 boxes respectively – the lists that describe them are extensive. The Print Journalism list's length blew out to 186 pages because of the number of items it contained; the Early Years list is 101 pages because of the complexity of the contents of its 281 files and the high research value of the material.

Millicent Weber, the lead author of this article, wrote a 117-page list for the contents of the 46 units in the Women and Literature series (2014.0047) and a smaller list for Correspondence with Libraries (2017.0004).

The box lists for the biggest series are truly monumental pieces of writing. Archivist Lachlan Glanville's finding aid for the 120-box General Correspondence Series (2014.0042) runs to 372 pages, while archivist Sarah Brown wrote 196 pages of metadata to describe the contents of 54 boxes in the Major Works series (2014.0045). Kate Hodgetts' main list for 184 records in Audio Recordings Produced and Received by Greer (2014.0040) is 50 pages but this is merely a guide to the detailed, multi-page, time-coded summaries Hodgetts has written for each audio file.

For a researcher, approaching and then navigating the paper behemoth that is the Greer Archive is a daunting task. Because the finding aids are digital documents, keyword searching offers an immediate 'in' for the time-poor researcher, a chance to identify archival records of direct and particular relevance to their topic of research. This kind of technique enables the researcher to check something specific; to locate already-familiar information and archival materials. But it does not offer a sense of the archive as a whole – of the trends, themes, hotspots and scandals that characterise it. By contrast, the work of the archivist – in the case of this article, the five months Buchanan spent creating the finding aid – enables rich insight into these broad, contextualising trends and themes. Here we argue that this broad, contextualising information – as well as the specific, deliberate detail – is present in the collection's metadata, and can be thematically interrogated using methods like sentiment analysis.

Sentiment analysis is a form of data mining: a means of enabling researchers to deal with unwieldy datasets. The term 'big data' describes these datasets: datasets that are overwhelming enough to disrupt the kinds of analysis traditionally used to understand them. A big dataset might be overwhelming because of the quantity of data it contains, or because those data are heterogeneous, unstable, unreliable or unstructured.[13] This is the kind of situation in which data mining comes into play.

Data mining refers to the use of computational techniques and programs to find patterns and trends in datasets. In the case of textual data, it might be used to identify words that appear often, that appear together often or that fluctuate in frequency. It can be used to look for words or phrases with particular characteristics – words of certain lengths, or particular parts of speech. It's been used for pretty much everything – from literary analysis of Shakespeare's plays,[14] to studies of the way that class is represented in American restaurant menus.[15]

The detailed finding aids that the Greer archivists have created do not only enable keyword searching. While more medium sized than strictly 'big data', they are nevertheless open to the possibilities of these new forms of interrogation. They are digital, richly textual and extensive. They also consist of a lengthy sequence of detailed descriptions of each file within the collection. In the case of the print journalism series, these 1277 descriptions run to 50,000 words of text, and make up the bulk of this massive finding aid. These are descriptions, too, that we know are comparable with one another – written by the same research archivist, following the same guidelines and using a consistent format and standardised language. Together, these features mean that they are able to be 'read' using computational techniques for data mining.

Sentiment analysis

Many different digital humanities techniques can offer productive insights into this kind of dataset. One of the simplest and most often-used examples is word frequency analysis. Figure 1 is an example of a word cloud created from the Print Journalism metadata, displaying the most frequent 100 words from the finding aid to this series, minus the terms 'contains', 'pages' and 'subjects', standard features of each series listing for the Greer archive.[16] This technique, although simple, offers some insight into the data's broad trends. It operates as a visually striking representation of the immediacy of data mining: one of its key appealing features for the time-poor researcher. Despite these affordances, word clouds and similar word frequency tools tend to flatten complex, granular datasets like archival metadata into homogenous corpora, and are also difficult to interrogate further: there is no obvious way to navigate, for example, between frequent keywords like 'Gardening' or 'Shakespeare' in this word cloud to the relevant descriptive metadata, let alone the relevant archival material. So, bearing this in mind, can approaches to textual data driven by technology lead to deeper insights, or insights that are more targeted to the needs of archival research?

Sentiment analysis is a process that evaluates and classifies text as positive or negative based on the strength of sentiment that it expresses. Unlike word frequency analysis, it preserves the existing structure of metadata while adding this additional layer of evaluative information. It therefore has potential as a helpful tool for researchers to locate archival material with strong affective content, while also posing questions, explored in detail later in this article, about the archivist's process in relation to objectivity and transparency. The sentiment analysis in this article was conducted
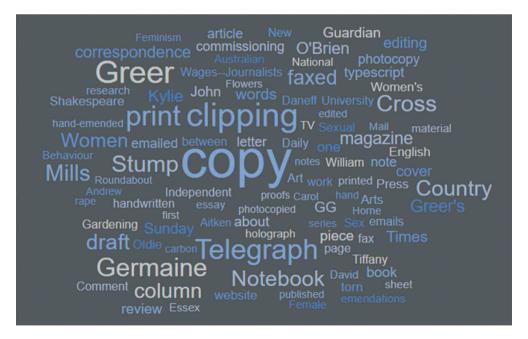

Figure 1. Word cloud from print journalism metadata. Displays most frequent 100 words, without 'contains', 'pages' and 'subjects'.

with the free program SentiStrength.[17] SentiStrength, which, although an imperfect tool, has been shown to encode sentiment with approximately the same degree of accuracy as human coders,[18] uses a lexical list to score each word in a text on a scale from −5 to 5, with neutral words scored at 0, strongly positive words scored at 5, and strongly negative words scored at −5. In the standard lexicon 'Betray' has a score of −3, and 'bigot' a score of −2. Their near neighbour, the acronym 'bff' (a colloquial, shortened form of 'best friends forever'), has a score of 4. Its inclusion in the vocabulary is a testament to the origins of sentiment analysis in making sense of social media data. SentiStrength looks at structure, too – it modifies scores based on the presence of words like 'not', 'very' and 'much', as well as the use of exclamation marks (and some emojis). It then gives each text in a dataset an overall positive and an overall negative score, calculated as the strongest positive word 'plus 1', and the strongest negative word 'minus 1'. The authors chose to use SentiStrength and its inbuilt lexicon because these are widely used in digital humanities research, making the results of this analysis potentially comparable to that of other textual corpora.

Weber ran the SentiStrength program across the descriptions from Buchanan's finding aid for print journalism. SentiStrength categorised 427 of these descriptions – almost exactly a third – as neutral, expressing no negative or positive sentiment. More than half the descriptions, 717, contained some kind of negative sentiment. The highly emotive nature of Greer's work – and the indignant anger of radical second-wave feminism – are too big to be contained, even by archival metadata. One fifth of the descriptions, 252, contained some kind of positive sentiment. Exactly half of these 252 were ambivalent, demonstrating both positive and negative sentiments concurrently.

This sentiment analysis revealed that the most strongly ambivalent file in the collection – with a negative score of −5 and a positive score of 4 – relates to the murder of James Bulger. The description for this file reads:

> Contains: clipping of column 'Love, not hate, will heal the tragedy of three small boys' 29 November 1993; draft print out, 3 pages, and fax cover sheet to Alan Rusbridger. Subjects: James Bulger murder; Murderers; Childhood; Humanity.

As a human reader, the reason for this ambivalence is strikingly clear. There is a clear contrast between the tragedy of murder, of this specific, appalling situation, and the stated message of the article, which is one of love, healing and redemption.

The most positive words used in the print journalism series, all with scores of '+3', are 'lover', 'romantic' and 'paradise'. These occur specifically in the context of the book title *Lady Chatterley's Lover*, in a mention of the TV Show *Desperate Romantics*, and subsequently a discussion of Romantic poets, and in a description of Melbourne as a 'white Anglo-Saxon-protestant paradise'. There is also an instance of the word 'romancing' in a column titled 'Germaine Greer is romancing the rainforest', a mention of another DH Lawrence novel, this time *Sons and Lovers*, and a perfect example that shows the pitfalls of computational readings of emotion – the National Institute for Clinical Excellence, which scores +3 for its excellence and an additional +1 for its acronym, NICE.

The strongly positive pieces of this archive are those that relate, explicitly, to romance and love – via the cultural conduits of DH Lawrence and *Desperate Romantics*, and

through Greer's love for the rainforest. These trends are crucial reminders of Greer's passion for literature, for popular culture and for the environment – passions that have recurred throughout her life's work.

And what of those 717 negative texts? In an archive like Greer's, it probably comes as no surprise that rape, abuse, domestic violence, death and women's health issues front the negative texts. A file relating to 'Seduction is a four-letter word', an article Greer wrote on rape for *Playboy*, is an important standout among these negative texts. It is there for the frequency of the terms 'rape' and 'rapist' in the description, but also for the word 'fuck': SentiStrength codes profanity with a score of minus two. Other profanities also populate the ranks of these highly negative texts. Anyone familiar with Greer's involvement with the explicit *Suck Magazine* in the early 1970s could probably hazard a guess what direction these contributions take.

Human coding

In her field-defining 1996 essay, Sue McKemmish analyses personal recordkeeping as an accumulation of 'evidence of me' – all the many versions of ourselves created via our relationships with others and society.[19] The Print Journalism series is rich with evidence of all the many journalistic selves Greer has experimented with: gonzo sex writer; theatre reviewer; satirist; gardening writer (for English- and German-language papers); feature writer; book reviewer; foreign correspondent; essayist; and columnist for the *Sunday Times*, the *Guardian*, the *Daily Telegraph*, the *Independent* and more. It also contains valuable evidence of the business of writing, the Greer who strove for the highest possible pay rate for her work and fought, always, for a fee when an article was not published.

However, Weber's use of the most basic of digital humanities techniques – the construction of a word cloud (Figure 1, above) – reveals that the list also contains evidence of another 'me', that of the archivist. In this particular instance, an archivist who worked as a print journalist for 25 years and who has written a book about the collapse of newspaper manufacturing and who has made an artist newspaper about dead newspapers.

Weber's word cloud demonstrates how strongly an insider, trade knowledge of print journalism has shaped the metadata. Many of the most common terms are trade ones: copy, clipping, typescript, article, editing, emails, handwritten, proofs. Following Greer's descriptions in a 1997 inventory she created to help prepare the archive for sale, the Print Journalist list focuses as much on the format of records as the content of the articles.[20] The Print series list even contains obsolete words like 'dinkus' (the newspaper term to describe a pre-formatted photo by-line for a column or columnist) and 'single column drop', terms that floated up from Buchanan's distant past as she read these records. The list also maps out shifts in technology from hot metal to computer typesetting, from cut-and-paste emendations (with scissors and Sellotape or glue) to cut-and-paste commands on a computer, from faxed proofs to emailed ones, from publication on newsprint to publication online.

Buchanan's history and politics has ensured that the Print Journalism finding aid is also 'evidence of us' – the 'us' being print journalists, especially print journalists in the United Kingdom. Buchanan listed the name of every newspaper and magazine editor

who commissioned work from Greer as well as the names of sub-editors who worked with her copy. The series contains evidence about pay rates and kill fees (the money paid to a writer when a commissioned piece does not run, another trade term that worked for this series). Freelancers normally keep this information to themselves but the print series lays it all out. The subject heading Buchanan created for this is 'Wages – Journalists' and it's one of the terms in Weber's word cloud. Copyediting is another common subject heading in the list to describe work that has all but disappeared now as newspapers continue to contract or move online and subeditors are offshored and then laid off.

The most common words in the list that describe content, rather than format, include Shakespeare, sex, flowers, women and gardening. There are probably some columns that touch on all five of these themes.

Print metadata: three case studies

In 1993 Greer began writing a regular column for the *Guardian*. The fourth of these was an extraordinary piece, a plea for mercy for the 11-year-old boys who murdered toddler James Bulger. The subject headings for this piece – 'James Bulger murder; Murderers; Childhood; Humanity' – are inadequate to describe the content of the column. One of the column's key themes is the challenge of being a child, and the number of people in Western societies who do not like children. These are arguments Greer developed in *Sex and Destiny* (1984).[21] The column expresses empathy for all children, including the two murderers and their victim, and lacerates adults who treat children poorly, especially mothers. It urges readers to try and imagine what it was like to be one of these boys, 'inattentive, unmotivated', 'collecting abuse and insult from people close and distant', and to use this exercise in empathy to see the murderers as human beings, not evil aliens. The complexity and the ambivalence of opinion pieces like this defy categorisation in straightforward subject headings.

Twelve months after the Bulger piece, Greer wrote another article with a similar plea for mercy for a murder. This time, the subject was Myra Hindley, who, with Ian Brady, murdered five children aged between 10 and 17.[22] Thirty years after committing these crimes, Hindley had agreed to an interview with the *Sunday Times* and told the journalist she wanted to be released. The strapline for Greer's column read: 'The hostility aroused by Myra Hindley's plea for release from prison says more about public vindictiveness than her fitness for freedom.'

Greer wrote: 'It is extraordinary that, knowing what she does about her role in the tabloid imagination, Myra Hindley should have the courage to beg for freedom.' The rest of the column was typically dense with ideas that defy summation. One theme common to these opinion pieces is Greer's strong contrarian impulse – if everyone else thinks something, she will think the opposite.

In an unusual move, Greer had filed some letters with the Hindley column. Ordinarily, her system of arrangement is such that the columns were filed in the Print series and the letters in response – notably the thousands of letters written by readers of her *Country Notebook* column in the *Daily Telegraph* (1999–2005) – were put in the General Correspondence series. Cataloguing the letters in response to the Myra

Hindley article involved noting down the names of each correspondent while considering what thematic detail to input in the metadata. At the back of the file, there were several letters from a David Astor (a lawyer) on thick, cream, watermarked paper – Greer had replied to one of these but could not find her letter – and 14 other letters from readers, all marked 'No action' at the top in black biro in Greer's hand. Beneath these, a letter from Hindley herself.

The letter from Hindley had also been marked 'no action' but as she read this letter Buchanan reflected that in keeping the letter Germaine Greer had taken action and that action had resulted in an immensely challenging moment for the archivist. The archivist's background as a print journalist shaped the metadata but the archivist's life as a mother of three daughters – all of whom were aged between 10 and 15, the same age as Hindley's victims – did too. The archivist's hand ran over the neat little words written by a woman who had sexually assaulted and murdered five children. The intimacy of this was horrifying; the only refuge was names and subject headings, the more objective the better.

After Buchanan listed the article, she washed her hands and sprinkled water over her face as a form of blessing. Alternate words for this list could be: horror, Greer Archive, the recordkeeping impulse, psychopaths, filing systems, personality disorders.

In 2006 Greer wrote an essay, again in the *Guardian*, about the death of Australian celebrity Steve Irwin. 'The animal world has finally taken its revenge on Steve Irwin. Germaine Greer on Australia's Diana moment' was the strapline on the front of the G2 section on 5 September.

The article was immediately syndicated and published, on 6 September, in the *Age* and the *Sydney Morning Herald*. Greer's agents, Gillon Aitken, were deluged with hate mail. As with the Myra Hindley column, Greer took the unusual step of filing letters, cards and printouts of emails with a clipping of her column. These responses fill two bulging folders. Although Greer had written many columns where her views were contentious and strongly opposed to the mainstream thought, none attracted this sort of response. The volume of responses and their ferocity is linked with email as a technology, particularly its speed and anonymity. It was made for vilification. The depth of misogyny in these files foreshadows the sort of trolling that younger feminists like Clementine Ford endure in the era of social media – and indeed the abuse that all women now routinely face online.[23]

Buchanan was cataloguing this material on the day that Donald Trump was unexpectedly beating Hillary Clinton in the US presidential race. Buchanan used two subject headings that expressed her judgement on the content of the files: 'Misogyny' and 'Ageism'. But she also tried to capture what appeared to be the legitimate criticisms of Greer's piece – that it had been written too soon and showed no sympathy for Irwin's grieving family. The subject headings were 'Grief' and 'Loss'.

Conclusion

The metadata for each of the specific files discussed in this article – the plea for compassion for James Bulger's murderers, the file of correspondence relating to Myra Hindley, and the hate mail received following Greer's article about Steve Irwin's death

– were picked up by the SentiStrength program as containing highly emotive content. The Myra Hindley file registered with a default positive score of 1, and a negative score of −4. The James Bulger file was the most ambivalent in the series, with the language used to describe it receiving scores of 4 and −5. The Steve Irwin material is, as Buchanan notes above, housed across two files: the file containing the article itself contains strongly ambivalent emotive language, scored at 3 and −4, while the metadata for the file containing the correspondence is more clearly negative, scoring at 1 and −4. The prevalence of contradictory, often dialectic arguments in Greer's opinion pieces frequently confounds the application of simplifying subject headings during the cataloguing process. But the process of sentiment analysis was designed with the internal inconsistencies of emotive content in mind. The prevalence among this listing of ambivalent metadata – accounting, as discussed above, for 126 of the descriptions, or about 10% of the finding aid – suggests that, despite the fears around homogenisation and simplification produced by the cataloguing process, the complexity of the collection is in fact discernible from its metadata.

As discussed above, there will necessarily be well-merited hesitations in relation to the use of computational programs to pick out something as human as emotive content, as shown in the example of the National Institute for Clinical Excellence. There are several difficult balancing acts that need to be achieved in cataloguing data, and each of these map to similar observations about the process of sentiment analysis.

The first of these balancing acts relates to the problem of historical language. For example, when *The Female Eunuch* was written, gender was overwhelmingly understood as something binary (woman, man), while more contemporary understandings describe gender as a spectrum. Likewise, in 1970, Greer and her contemporaries were agitating for women's liberation; the word feminism came later. Throughout the process of cataloguing this vast archive, the Greer archivists developed a shared thesaurus, a standardised wordlist combining Library of Congress subject headings and selected additions where necessary. One of the goals that shaped the approach taken to this collection was to catalogue the Greer Archive so that it would be as findable and as searchable as possible. This meant the use of standardised headings was essential. Another goal was to catalogue the archive in a manner that was as sensitive as possible to the different communities and voices represented and discussed in it. Political sensitivity and awareness in the twenty-first century are somewhat at odds with the use of a thesaurus with roots in the nineteenth century.[24]

A similar difficulty presented itself in the use of the SentiStrength vocabulary. After an initial pass of the data, several anomalies specific to the Greer collection became apparent. The word 'cross' – a frequent occurrence, because of Greer's property at Stump Cross and her publisher Stump Cross Books – has a negative score of −1. Ironically, given the context of this particular application of sentiment analysis, the word 'misogyny' scores as neutral. Despite the temptation of amending the lexicon to better reflect these elements of the collection, we hesitated to make these changes. Again, this was because of the promises of standardisation and comparability. The SentiStrength lexicon is widely used across the digital humanities. Keeping it intact means that the sentiment contained in this metadata remains comparable to that of other textual corpora, developing the potential for future research to examine the sentiment of this archival metadata in a broader social or cultural context.

The final balancing act involves the push and pull between objectivity and affective content, a tension at the core of cataloguing a collection like this one. Sentiment analysis sits directly across this seeming binary. On the one hand, it is a digital process – it is a method that counts words, crunches numbers and processes vast amounts of data in seconds. But on the other, it is a highly qualitative one. It is a way of picking out the human traces, the emotional qualities that hide in these intimidating datasets. There is a minimalist approach, a kind of objectivity, which many of us seek to achieve as cataloguers. But each of us who have worked with it can attest that Greer's archive is drenched in affect. The fact that this can be read and tracked by a computer program speaks to the impact that this kind of content has on its cataloguers. And it also suggests that metadata reveals, rather than obscures, the collection's emotive landscape.

Notes

1. Germaine Greer, *The Female Eunuch*, MacGibbon & Kee, London, 1970.
2. Rachel Buchanan, 'The Iran Album (1974): Some Sleeve Notes', *Archivaria*, vol. 84, Fall 2018, pp. 40–75.
3. Carolyn Pedwall and Anne Whitehead, 'Affecting Feminism: Questions of Feeling in Feminist Theory', *Feminist Theory*, vol. 13, no. 2, 2012, pp. 115–29; Michelle Caswell and Marika Cifor, 'From Human Rights to Feminist Ethics: Radical Empathy in Archives', *Archivaria*, vol. 81, Spring 2016, pp. 23–43.
4. Eira Tansey, 'Archives Without Archivists', *Reconstruction: Studies in Contemporary Culture*, vol. 16, no. 1, 2016, available at <https://web.archive.org/web/20161031084443/http://reconstruction.eserver.org:80/Issues/161/Tansey.shtml>, accessed 6 January 2018.
5. ibid.
6. Mark Greene and Dennis Meissner, 'More Product, Less Process: Revamping Traditional Archival Processing', *The American Archivist*, vol. 68, no. 2, 2005, pp. 208–63.
7. MS9982; finding aid available at <https://nla.gov.au/nla.cat-vn3789379>, accessed 6 January 2019.
8. MS 3665; and MS 7599; finding aids available at <https://nla.gov.au/nla.cat-vn233595> and <https://nla.gov.au/nla.cat-vn514768>, accessed 6 January 2019.
9. Graeme Powell, 'The Collecting of Personal and Private Papers in Australia', *Archives and Manuscripts*, vol. 24, no. 1, 1996, pp. 62–77.
10. Rachel Buchanan, 'Foreign Correspondence: Journalism in the Germaine Greer Archive', *Archives and Manuscripts*, vol. 46, no. 1, 2018, pp. 18–39.
11. For example, description of archives in the Australian context, especially government archives, tends to focus on the description of provenance entities and relationships and functional classification of records (including the application of functional thesauruses).
12. Chris Hurley, 'The Making and Keeping of Records: (2) The Tyranny of Listing', *Archives and Manuscripts*, vol. 28, no. 1, 2000, p. 9; Chris Hurley, 'The Making and Keeping of Records: (1) What Are Finding Aids for?', *Archives and Manuscripts*, vol. 26, no. 1, 1998, pp. 57–77.
13. Amir Gandomi and Murtaza Haider, 'Beyond the Hype: Big Data Concepts, Methods, and Analytics', *International Journal of Information Management*, vol. 35, no. 2, 2015, pp. 137–44.
14. Jonathan Hope and Michael Witmore, 'The Very Large Textual Object: A Prosthetic Reading of Shakespeare', *Early Modern Literary Studies*, vol. 9, no. 3, 2004, pp. 1–36.
15. Dan Jurafsky, Victor Chahuneau, Bryan R. Routledge and Noah A. Smith, 'Linguistic Markers of Status in Food Culture: Bourdieu's Distinction in a Menu Corpus', *Journal of Cultural Analytics*, 18 October 2016, https://doi.org/10.22148/16.007

16. Word pairs 'Greer' and 'Greer's' and 'women' and 'women's' appear in the word cloud. The decision not to stem this data derives from the semantic differences between these terms: the fact that the metadata not only discusses Greer and women but strongly reiterates their ownership of materials, concepts, political movements etc. contained in the archive becomes a political statement when taken in light of the gendered history of ownership.

17. The software program is available at <http://sentistrength.wlv.ac.uk/>, accessed 9 January 2019.

18. Mike Thelwall, 'The Heart and Soul of the Web? Sentiment Strength Detection in the Social Web with SentiStrength', in Janusz Holyst (ed.), *Cyberemotions: Collective Emotions in Cyberspace*, Springer, Berlin, 2017, pp. 119–34. For evaluation of the program's usefulness also see Beth Driscoll, 'Sentiment Analysis and the Literary Festival Audience', *Continuum*, vol. 29, no. 6, 2015, pp. 861–73.

19. Sue McKemmish, 'Evidence of Me', *The Australian Library Journal*, vol. 45, no. 3, 1996, pp. 174–87.

20. Buchanan, p. 20.

21. Germaine Greer, *Sex and Destiny: The Politics of Human Fertility*, Secker & Warburg, London, 1984.

22. Germaine Greer, 'Revenger's Tragedy (Myra Hindley)', *The Guardian*, 20 December 1994, University of Melbourne Archives, Germaine Greer Archive, 2014.0046.00391.

23. Tania Levey, *Sexual Harassment Online: Shaming and Silencing Women in the Digital Age*, Lynne Rienner Publishers, Boulder CO, 2018; Laurie Penny, 'Who Does She Think She Is?', *Longreads*, 28 March 2018, available at <https://longreads.com/2018/03/28/who-does-she-think-she-is/>

24. Karen Nuckolls, 'Subject Access to Diversity Materials: The Library of Congress Subject Heading Shortfall', *The Reference Librarian*, vol. 21, nos. 45–46, 1994, pp. 241–51. Also see William E Studwell, *Library of Congress Subject Headings: Philosophy, Practice and Prospects*, Routledge, New York, 1990.

Disclosure statement

Notes on contributors

Millicent Weber is a lecturer in English at the Australian National University, and has also worked at the University of Melbourne Archives and National Library of Australia. She researches the intersections between live and digital literary culture, and is the author of *Literary Festivals and Contemporary Book Culture* (Palgrave Macmillan, 2018). She has published articles in peer-reviewed journals including *Continuum* and *Convergence* and literary journal *Overland*, and co-edited *Publishing Means Business: Australian Perspectives* (Monash University Publishing, 2017).

Rachel Buchanan's most recent book, *Ko Taranaki Te Maunga* (BWB Texts, 2018), is about shame, families and recordkeeping as a powerful form of non-violent resistance to colonisation. Between October 2015 and March 2018, Buchanan was curator, Greer Archive, University of Melbourne Archives and she has published widely on the archive and on her work with the Greer team. *Stop Press* (Scribe, 2013) and *Melbourne Sirius*, an artist newspaper funded by a State Library of Victoria creative fellowship (2014), are two records of her first career as a newspaper journalist.