Seams and edges: Dreams of aggregation, access & discovery in a broken world

Abstract

Visions of technological utopia often portray an increasingly 'seamless' world, where technology integrates experience across space and time. Edges are blurred as we move easily between devices and contexts, between the digital and the physical.

But Mark Weiser, one of the pioneers of ubiquitous computing, questioned the idea of seamlessness, arguing instead for 'beautiful seams' — exposed edges that encouraged questions and the exploration of connections and meanings.

With discovery services and software vendors still promoting 'seamless discovery' as one of their major selling points, it seems the value of seams and edges requires further discussion. As we imagine the future of a service such as Trove, how do we balance the benefits of consistency, coordination and centralisation against the reality of a fragmented, unequal, and fundamentally broken world.

This paper will examine the rhetoric of 'seamlessness' in the world of discovery services, focusing in particular on the possibilities and problems facing Trove. By analysing both the literature around discovery, and the data about user behaviours currently available through Trove, I intend to expose the edges of meaning-making and explore the role of technology in both inhibiting and enriching experience.

How does our dream of comprehensiveness mask the biases in our collections? How do new tools for visualisation reinforce the invisibility of the missing and excluded? How do the assumptions of 'access' direct attention away from practical barriers to participation?

How does the very idea of systems and services, of complex and powerful 'machines' ready to do our bidding, discourage us from seeing the many, fragile acts of collaboration, connection, interpretation, and repair that hold these systems together?

Trove is an aggregator and a community; a collection of metadata and a platform for engagement. But as we imagine its future, how do avoid the rhetoric of technological power, and expose its seams and edges to scrutiny.

Paper

In March 1930 the Sydney Electrical and Radio Exhibition opened in a blaze of excitement. Aboard his yacht in Genoa, inventor Guglielmo Marconi triggered a radio signal that reached across the world and switched on more than 2800 electric lights at the Sydney Town Hall. 'All in less than a second!', exclaimed the *Sydney Mail*, 'Here was magic! Arabian nights recede into remoteness: their magic was nothing compared to this'.[1]

Radio had 'eliminated time and distance', argued the *Sydney Morning Herald*, seeing in the exhibition a future where electricity would free the world from drudgery.[2] About a month later the British and Australian Prime Ministers spoke for the first time via wireless telephone. The British PM, Ramsay McDonald, suggested that the technology 'would be the means of knitting the two countries closer and closer together'. 'These were days for the annihilation of time and space', he proclaimed.[3]

From railways to the telegraph, radio, and the internet, the progress of technology has often been imagined as a battle against time and space. Progress has been measured in the seconds we save, in the distances we conquer, in the barriers of terrain and politics we bridge.

In the realm of information this march of conquest is accompanied by adjectives such as 'instantaneous' and 'seamless'. No need to wend your way between separate sources and services, technology promises a future beyond silos.

You don't have to look too hard to find software and service vendors touting the promise of 'seamless discovery'. Indeed, it turns out that 'Seamless Discovery' itself is the registered trademark of a video discovery platform used by Foxtel and others.[4]

In the library world, seamless discovery is commonly associated with what are variously called 'next-generation catalogues', 'web-scale discovery services' or 'discovery layers'.[5] The idea is familiar and seductive. Instead of forcing searchers to construct multiple queries across a variety of databases, systems and interfaces, these services aggregate metadata from different sources and offer access through a single search portal. The march of library technology promises to annihilate the legal and technological barriers that interrupt our information-seeking journey. A seam-free service is one that maximises ease-of-use.

Library users already have a very clear picture of what such a service might look like. Every day they undertake a wide variety of social and economic exchanges mediated through the infrastructure of search. Google might not be the only platform for online discovery, but it has played a central role in re-engineering our understanding and expectations of online experience. Search is no longer just a task to be accomplished in pursuit of a particular goal — to find a desired resource or piece of information. Ours is increasingly a 'culture of search' where the technologies of discovery are naturalised 'into the backgrounds, fabrics, spaces and places of everyday life'.[6] I search, therefore I am.

It's natural then that users of other discovery services will approach them with a set of expectations shaped by the Googlisation of modern culture. It's not just the simplicity of that single search box, it's our faith that search will *just work*. Every time Google

responds to our query about some obscure piece of television trivia with 152 million results, we cannot fail to be impressed by the power at our fingertips. Every time Google predicts our query or customises our results we are beset with awe — a combination of fear and wonder. This must be magic.[7]

Library services cannot compete with Google's oracular power, but they can at least aim to offer users a comparable level of simplicity. The features of 'next-generation catalogues' or discovery layers tend to follow a familiar check-list: single search box, faceted navigation, and relevance-ranked results. The pursuit of seamless discovery likewise mirrors Google's totalising reach. One search box to access a whole world of data.

There's nothing wrong with this — we all want to make life as easy as possible for the people who use our services. The question is how the pursuit of a Google-like experience constrains our options and assumptions. Despite the mathematical foundations of Google's PageRank algorithm there are politics at work in calculations of relevance and criteria for inclusion.[8] Google's dominance gives it immense power in presenting to us an image of the world constructed to it's own secret formula. This power bears ontological weight — if we can't find something on Google does it exist? If we are concerned with absence as well as inclusion, with addressing the silences within our cultural record, we need to wary of sharing in Google's aura of completeness.

Seams are not simply obstacles to a smooth user experience, they're reminders that our online services are themselves constructed. There's nothing natural or inevitable about a list of search results. Mark Weiser, one of the pioneers of ubiquitous computing, argued against seamlessness because it made everything seem the same. Instead he imagined systems with 'beautiful seams'.[9] The possibilities of 'seamful design' have been taken up by other researchers, exploring ways that users can be empowered to discover and manipulate their contexts and connections.[10]

As Mitchell Whitelaw notes 'seamfulness is also an ethical and political stance' — it's a commitment to exposing the interpretative distance between our collection data and its online representation.[11] There are opportunities here not only for transparency, but to explore alternatives to Google's template for discovery. Research into the visualisation of large cultural heritage collections, by Whitelaw and others, has emphasised that search is only one way of representing a collection.[12] By focusing on the stylish minimalism of the search box, we discard opportunities for traversing relationships, for fostering serendipity, for seeing the big picture.

It's important to recognise, however, that this type of research is not aimed at supplanting search, nor building a better Google. Nor indeed should alternative collection interfaces be judged on narrow measures of utility. This is *building as critique* — each alternative interface offers a means of questioning our assumptions about the discovery of online collections. As Matt Ratto argues in his discussion of 'critical making', 'these material interventions provide insubstantiations of how the relationship between society and technology might be otherwise constructed'.[13] By playing around with our expectations we can start to think differently, to develop new metaphors for our online experience.

My own *Eyes on the past*, which allows you to find your way into Trove's digitised newspapers through machine recognised faces and eyes, is far from a practical

discovery tool.[14] But building on my earlier work using facial detection technology as a means of archival intervention, it opens up questions about the lives embedded within our collections — we see them differently, we feel differently.

A Google-like search experience offers utility at the expense of critique. Its technologies are black boxed, its assumptions obscured. How do those of us in the discovery business respond? How do we create a buffer for critical reflection while still meeting user expectations? By unpicking a few seams, cultural institutions can open up a space for discussion, but what does this actually mean for a service such as Trove that must deal with thousands of users a day?

I'd suggest we start with an acknowledgement of our limits, an attempt to trace the edges and the fractures that are too often glossed over in our pursuit of seamlessness. I also think we should take our metaphors seriously, not just as marketing hype, but as the means by which structure the realm of what is possible. Let's start by admitting what Trove is not:

1. Trove is not perfect

2. Trove is not everything

3. Trove is not a machine

Trove is not perfect

Trove is an aggregator. It pulls together metadata from a variety of different sources, applies some normalisation across the required fields, and sends the results off to be indexed. With close to 400 million resources harvested from hundreds of contributors through an assortment of different pipelines, it's inevitable that there will be errors and oddities. Descriptive standards vary, and sometimes the assumptions Trove makes about the data it's getting are wrong.

If you want to see errors, of course, you can head along to Trove newspapers zone where the limitations of Optical Character Recognition are on display for all to see. Unlike some full-text databases, Trove exposes the raw output of its OCR processing. The accuracy of OCR is heavily dependent on the quality of the source material which, in the case of historical newspapers, varies considerably.[15]

A few years ago, as part of separate research project, I made an attempt to estimate OCR accuracy in Trove across a sample of 10,000 newspaper articles.[16] I basically just compared the OCR output to a dictionary list of words and calculated the accuracy of each article as a percentage of the total number of words. Variations were considerable across both time and titles, but the average was around 85%. A much more rigorous analysis of the British Library's digitised 19th century newspapers found an overall word accuracy of 78%.[17]

Trove's transcriptions are improving all the time thanks to the efforts of thousands of online volunteers who correct the raw OCR output. Astonishingly, more than 130 million lines of text have been corrected by Trove users, in what is rightly touted as a highly successful crowdsourcing initiative. But it's also important to put this effort in

perspective. Head across to the Trove newspapers zone and enter 'has:corrections' into the search box to retrieve all the articles that have at least one crowdsourced correction.[18] At the time I wrote this, the figure was 5,273,600 or just 3.6% of the total number of newspaper articles in Trove. Paul Hagon's analysis of Trove crowdsourcing behaviour also indicates there is a flattening out of growth in corrections. Despite their important efforts, Trove's volunteers will never be able to produce a perfect rendering of the newspaper content.

But what is 'perfection' anyway? OCR accuracy is important only in so far as it supports the interests and activities of users. For the purposes of discovery the accuracy of common search terms such as names, places or events are likely to be most important. But a much broader range of words would be significant in an analysis of changes in language across time. Accuracy is something that need to be assessed and understood within the context of a specific research activity. Researchers using digitised text collections need to consider the impact of technologies such as OCR on their methodologies, or else, in Tim Hitchcock's words, 'This is roulette dressed up as scholarship'.[19]

Services like Trove can support rigorous digital scholarship by exposing as much information as possible about the technologies they employ and any known limitations. This applies not just to OCR, but to fundamental technologies such as keyword search and relevance ranking. If we are developing resources for scholarly use we cannot simply black box our tech and trade on trust. That's Google's game. We have to be prepared to expose configurations and assumptions so that analyses can be replicated and exposed to critique.

QueryPic is a simple tool that visualises search results in the Trove newspapers zone. QueryPic lets you see patterns and trends across the whole database but, as the help system warns, it creates 'sketches, not arguments' — critical interpretation is always required.[20]

When did the 'Great War' become the 'First World War'?[21] QueryPic can be used to explore this shift in terminology, but if you examine the results closely you'll notice a small bump in the graph indicating that the term 'World War I' was being used *during* World War I. Huh? If you drill down through the results you'll find that this is because Trove users have been busily adding the tag 'World War I' to selected articles, and by default Trove searches user tags and comments as well as article text. The bump is an artefact of Trove's search configuration.

Trove's primary function is discovery — to make it as easy as possible for people to find things they're interested in. But the sort of fuzziness that supports discovery works against other forms of analysis. We should make these sorts of assumptions more obvious, and provide opportunities for researchers to question things like relevance ranking. By showing our seams, exposing our imperfections, we have the opportunity to educate. As well as helping people use Trove, we can open up bigger questions about the way search works on the web.

Trove is not everything

There's nothing natural about our cultural collections or their digital representations — they have been created by many acts of selection, neglect, vision, accident and planning.

If you graph the number of newspaper articles in Trove by state and year you'll notice a

Trove newspaper articles by state
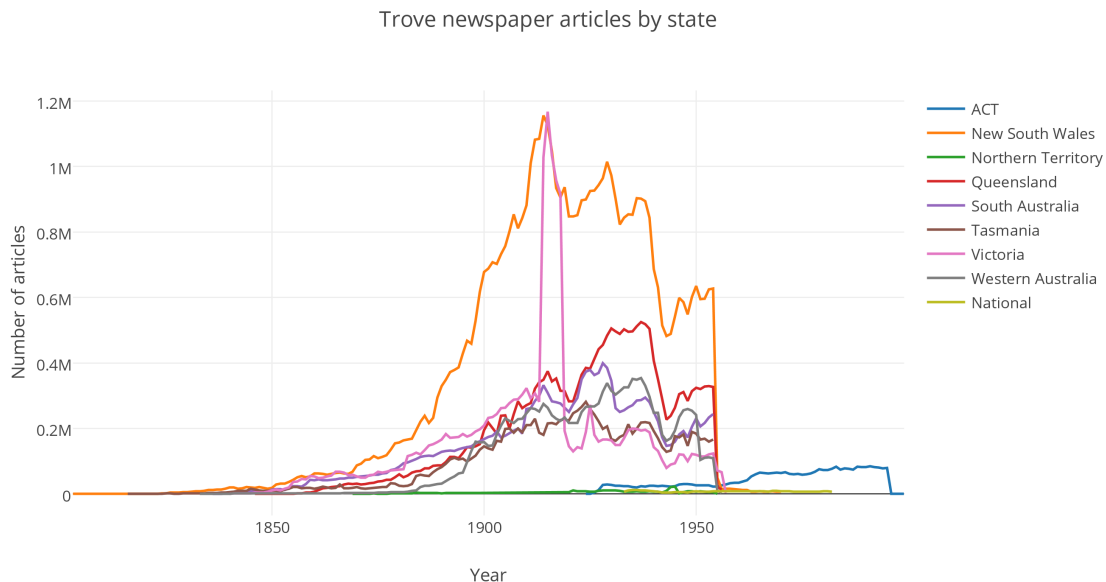
rather dramatic spike around 1914.[22]
Figure 1 — Trove newspaper articles by state/year

Why? Were more newspapers printed during the war era? The answer is simply funding. As part of the Australian Newspaper Digitisation Program, the NSW and Victorian State Libraries have chosen to invest in the digitisation of newspapers from the World War I period. The contents of Trove's newspaper zone, like any online collection, is constructed — shaped by many competing priorities. The consequences of this process are not always obvious.

In a competition for resources what gets digitised and why? There's a danger that the sheer scale of aggregation services like Trove will reinforce existing prejudices. People already struggling for visibility and recognition within our cultural record might be lost amidst the overwhelming numbers of the safe and the sanctioned. The ontological weight of search can too easily equate absence with non-existence.

But aggregation also offers new opportunities for analysis. Questions of representation and diversity can be explored through the metadata itself. Mitchell Whitelaw notes that some collection interfaces are already exploring ways of representing absence. Perhaps we can extend this evolving language across large aggregated collections to reveal not only what is found, but what is missing.
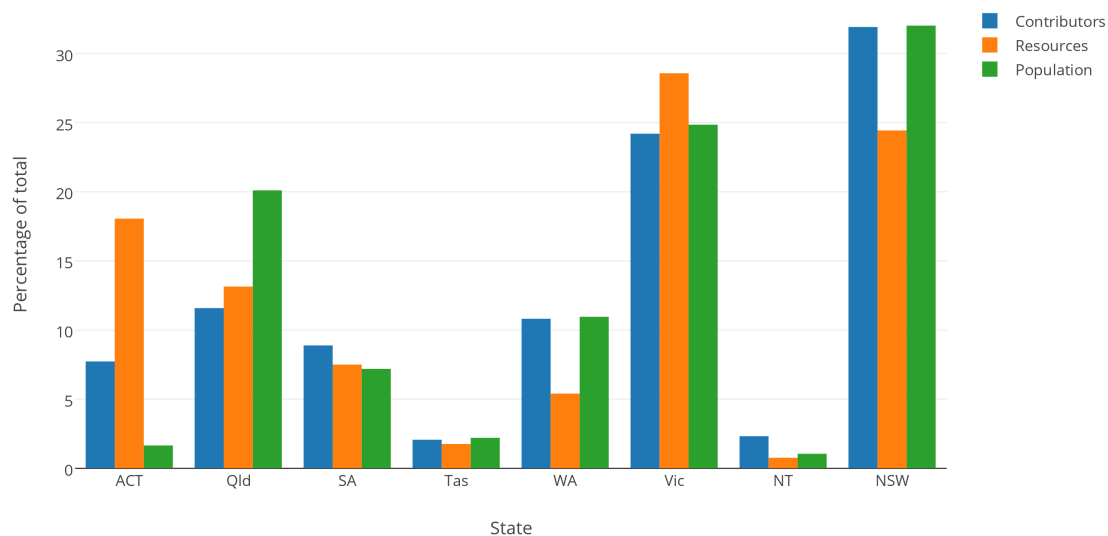
Figure 2 — Trove resources and contributors by state

By way of a quick example, I used the Trove API to harvest raw numbers of holdings and contributors for each state. It was a simple matter to combine these with population data to create a crude graph of resource representation by state.[23] Obvious anomalies, such as Queensland's apparent underrepresentation, might be simply explained by demographics, but the point is that aggregated data enables us to frame these sorts of questions without undertaking a major research project.

Perhaps more interestingly, I was able to easily compare the languages spoken at home in Australia, according to the 2011 Census, with the languages of resources in Trove's book zone.[24] It's fascinating to consider how we might use socio-economic data to slice our cultural collections across the grain to reveal different patterns of access and exclusion.

There are other opportunities as well. Like Trove, the Digital Public Library of America aggregates metadata from a wide range of cultural organisations. The DPLA has taken a public stance on diversity, monitoring its own holdings to highlight questions of underrepresentation, and working proactively to fill known gaps.[25] By admitting the constructed nature of our collections, the gaps and the silences as well as their strengths, perhaps aggregations like Trove can become sites of both analysis and activism.

Trove is not a machine

Trove is not a single application, it's a complex system with multiple components. This size and complexity focuses our attention on the technology — on the lines of code and racks of servers. But the system only exists to support human creativity and cooperation. Is it a machine, a community, or something else?

I often talk about Trove as a platform — it can be built upon in many ways, both through code and collaborations.[26] In particular, by providing an open API, Trove invites the public to create new tools, analyses and interfaces.

But there are metaphorical dangers lurking here as well. Social media services such as Facebook and YouTube also describe themselves as platforms — staking out a space alongside traditional media outlets, while seeking to expand through developers and new technology partners. As Tarleton Gillespie notes, 'these terms matter as much for what they hide as for what they reveal'.[27] In this case, the 'platform' label can divert attention away from the analysis of business models involving the monetisation of personal data.

If we are to embrace the 'platform' metaphor we must also be ready to unpack its implications. Writing about an earlier generation of information infrastructure metaphors — superhighways, virtual communities and digital libraries — Peter Lyman argued that such terms contained 'an indirect dialogue about questions of social and economic justice in an information society'.[28] If we want progressive platforms we need to honestly address issues of openness, participation, and accessibility. Every API is an argument and no data is ever truly 'open'.[29]

For me the term 'platform' speaks of something unfinished — an invitation and an opportunity. Trove is permanently under construction, constantly improved through the labours of its developers and community. This is most evident in the work of Trove's text correctors, whose many small acts of repair help the technology to function more efficiently. But each tag or comment also changes Trove — aiding discovery, adding context, or creating new connections.

The Trove API is not merely a plaything for tinkerers like me. No interface can ever serve the needs of all users — there will be inevitably be biases and assumptions that limit engagement. But the API at least keeps open the possibility of alternative Troves that address existing biases and meet the needs of specific communities in a way that a single, centralised portal can never do.

Other Trove-building activity is less visible, and the responsibilities more distributed. For example, Trove is currently working with Victorian Collections to bring many small, local collections from across Victoria into Trove.[30] But this collaboration is itself built on the labours of many people over many years — from the Museums Australia staff who train community groups, to the local volunteers who painstakingly digitise and describe their collections. Trove helps bring these efforts to the attention of the web, and is itself enriched.

As Peter Lyman notes, for all the new terms we have for systems and devices we have thus far failed to find a language to describe online collaboration and social engagement. Instead we fall back on the awful term 'user' — 'a word that places technique at the centre, and even contains a hint of dependence upon or subordination to technology'. By drawing attention away from 'the machine' to the many small acts that sustain and enlarge a service such as Trove, we create a space where language might evolve.

Broken worlds

Instead of visions of technological progress, Steven J. Jackson presents a vision of a fundamentally broken technosocial world barely held together by numerous acts of concern and repair.[31] Most technological futures are ultimately alienating and disempowering — we are passive consumers of the latest wonders and gadgets. By focusing on 'repair', as Jackson suggests, we see the human agency at work, the possibilities for change.

Similarly, by seeing our seams and edges as sites of repair rather than speed bumps in the onward march of progress, we can open spaces for dialogue, for sharing, and for learning — for imagining something different.

References

1. 'When Marconi Switched on the Lights The Sydney Electrical and Radio Exhibition', *Sydney Mail* (NSW), 2 April 1930, p. 20. <http://nla.gov.au/nla.news-article160633081> ↵

2. 'Tales of the Genii', *The Sydney Morning Herald* (NSW), 27 March 1930, p. 10. <http://nla.gov.au/nla.news-article16636946> ↵

3. 'Wireless Telephony. England and Australia. Prime Ministers Converse. Brisbane, April 30', *Cairns Post* (Qld.), 2 May 1930, p. 4. <http://nla.gov.au/nla.news-article41057074> ↵

4. Digitalsmiths Seamless Discovery, <http://www.digitalsmiths.com/solutions/seamless-discovery/features/> ↵

5. Joshua Barton and Lucas Mak, 'Old Hopes, New Possibilities: Next-Generation Catalogues and the Centralization of Access', *Library Trends*, vol. 61, no. 1, 2012, pp. 83–106. <http://muse.jhu.edu/journals/library_trends/v061/61.1.barton.html> ↵

6. Ken Hillis, Michael Petit, and Kylie Jarrett, *Google and the Culture of Search*, Routledge, 2013, p. 5. ↵

7. Ken Hillis, Michael Petit, and Kylie Jarrett, *Google and the Culture of Search*, Routledge, 2013, p. 14ff. ↵

8. Lucas D. Introna and Helen Nissenbaum, 'Shaping the Web: Why the politics of search engines matters', *The Information Society*, vol. 16, no. 3, 2000, pp. 169–185. <http://www.tandfonline.com/doi/abs/10.1080/01972240050133634>; Laura A. Granka, 'The Politics of Search: A Decade Retrospective', *The Information Society*, vol. 26, no. 5, 27 September 2010, pp. 364–374. <doi:10.1080/01972243.2010.511560> ↵

9. Quoted in Matthew Chalmers and Ian MacColl, 'Seamful and seamless design in ubiquitous computing', in *Workshop At the Crossroads: The Interaction of HCI and Systems Issues in UbiComp*, 2003. ↵

10. Matthew Chalmers and Areti Galani, 'Seamful interweaving: heterogeneity in the theory and design of interactive systems', in *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques*, ACM, 2004, pp. 243–252. http://dl.acm.org/citation.cfm?id=1013149 ↵

11. Mitchell Whitelaw, 'Representing Digital Collections', in *Performing Digital: Multiple Perspectives on a Living Archive*, ed. David Carlin and Laurene Vaughan, Ashgate Publishing, Farnham, UK, 2014. ↵

12. See for example the DL2014 Workshop, 'The Search Is Over! Exploring Cultural Collections with Visualization'. http://searchisover.org/ ↵

13.Matt Ratto, 'Critical Making', in *Open Design Now: Why Design Cannot Remain Exclusive*, ed. Bas van Abel, Lucas Evers, and Peter Troxler, BIS Publishers, Amsterdam, The Netherlands, 2011. http://opendesignnow.org/index.php/article/critical-making-matt-ratto/ ↵

14.*Eyes on the past*, <http://eyespast.herokuapp.com/>. For context see 'Eyes on the past', <https://storify.com/wragge/eyes-on-the-past>. ↵

15.Rose Holley, 'How Good Can It Get?: Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs', D-Lib Magazine, vol. 15, no. 3/4, March 2009. <http://www.dlib.org/dlib/march09/holley/03holley.html> ↵

16.Tim Sherratt, 'Mining for meanings', Harold White Fellowship public lecture, National Library of Australia, 8 May 2012. http://discontents.com.au/mining-for-meanings/ ↵

17.Simon Tanner, Trevor Muñoz, and Pich Hemy Ros, 'Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive', *D-Lib Magazine*, vol. 15, no. 7/8, July 2009. <http://www.dlib.org/dlib/july09/munoz/07munoz.html> ↵

18.<http://trove.nla.gov.au/newspaper/result?q=has%3Acorrections> ↵

19.Tim Hitchcock, 'Historyonics: Academic History Writing and its Disconnects'. <http://historyonics.blogspot.ca/2011/10/academic-history-writing-and-its.html>; Ian Milligan, 'Illusionary Order: Cautionary Notes for Online Newspapers', ActiveHistory.ca. <http://activehistory.ca/2012/03/illusionary-order/> ↵

20.<http://dhistory.org/querypic/help/> ↵

21.<<http://dhistory.org/querypic/43/>. See also Tim Sherratt, 'When did the "Great War" become the "First World War"?'. http://discontents.com.au/when-did-the-great-war-become-the-first-world-war/ ↵

22.TroveNewspapers ↵

23.<https://plot.ly/~wragge/19> ↵

24.<https://plot.ly/~wragge/21> ↵

25.'Digital Public Library of America » Blog Archive » Diversity and the DPLA'. <http://dp.la/info/2014/01/17/diversity-and-the-dpla/> ↵

26.Tim Sherratt, 'From portals to platforms: building new frameworks for user engagement', presented at the LIANZA 2013 Conference, Hamilton, New Zealand, 21 October 2013. <http://www.nla.gov.au/our-publications/staff-papers/from-portal-to-platform> ↵

27.Tarleton L. Gillespie, 'The Politics of "Platforms"', *New Media & Society*, vol. 12, no. 3, 1 May 2010. http://papers.ssrn.com/abstract=1601487 ↵

28.Peter Lyman, 'Information Superhighways, Virtual Communities and Digital Libraries: Information society metaphors as political rhetoric', in *Technological Visions: The Hopes and Fears that Shape New Technologies*, ed. Marita Sturken, Douglas Thomas, and Sandra J Ball Rokeach, Temple University Press, Philadelphia, 2004, pp. 201–218. ↵

29.Tim Sherratt, '"A map and some pins": Open data and unlimited horizons', presented at the Digisam conference on Open Heritage Data in the Nordic Region, Malmö, 25 April 2013. http://discontents.com.au/a-map-and-some-pins-open-data-and-unlimited-horizons/ ↵

30.'Growing together – Trove and Victorian Collections', <http://www.nla.gov.au/blogs/trove/2014/11/11/growing-together-trove-and-victorian-collections>. ↵

31.Steven J. Jackson, 'Rethinking repair', *Media meets technology*, MIT Press, 2013. ↵