Visualising Archival Collections - The Visible Archive Project

Dr Mitchell Whitelaw
Faculty of Arts and Design
University of Canberra

Our archives, libraries, and other cultural collections are increasingly digital collections. With the digitisation of historical material, and the collection of contemporary "born digital" records, archives and other such collections have become custodians not only of objects, but of data, and responsible for its structure, preservation, and presentation. In discussing the promises and perils of digitisation for cultural collections, Abby Smith vividly describes the "variability" and "plasticity" of data, and the challenges this poses for digital recordkeeping.[1] However these dynamic qualities of data also provide opportunities for new approaches that can help address some of the challenges of the digital.

We deal with visual representations of data on a daily basis; every web site and computer interface, not to mention spreadsheet chart or graph, is essentially data in visible form. The plasticity of data, its computability, means that data can be represented visually in any number of different forms, some familiar and some not. Given a certain body of data then, the question is how best to display or present it; an entire field of research and practice - *data visualisation* - is founded on this question.

This paper presents a research project that essentially addresses the same question, in the context of cultural collections. If we consider an archival collection - in this case the collection of the National Archives of Australia - as a rich set of data: how best to display it? Supported by the National Archives under its 2008 Ian Maclean award, the Visible Archive project developed new techniques for the interactive display of archival data. The work presented here shows that visualisation can play a significant role in supporting our undertanding of, and access to, large archival collections, at levels from the whole collection, to the individual item.

Aims and Context

As Friendly and Denis show, the visual display of quantitative information has a long history, spanning cartography and scientific diagrams, graphs and charts.[2] More recent visualisation is characterised by the development of interactive computing, and the ability to manipulate visual representations directly. Friendly and Denis also comment on the recent proliferation and interdisciplinarity of visualisation work. Friedman's 2007 survey of data visualisation provides a sense of this diversity, as well as the dominant influence of networked data sources and presentation techniques.[3] Current visualisation practice is a broad and growing field spanning information technology, the digital humanities, design, and art; this interdisciplinarity informs the background and context of this project.

In tandem with this broadening of visualisation comes a growing recognition that science is not alone in generating ever-increasing volumes of data, or in needing to access and interpret that data effectively. Studies of history, society and culture make increasing use of digital materials and methodologies, including visualisation.[4, 5] Researchers in the field have begun to recognise the potential of visualisation; Lev Manovich for example describes research into "visualising cultural patterns."[6] Examples of visualisations of large collections are scarce, but include George Legrady's 2005 *Making Visible the Invisible*, a dynamic visualisation of activity in the collection of the Seattle Central Library.[7] Jeanne Kramer-Smyth's ArchivesZ project is more relevant to this project - an interactive tool using visualisation to support search and exploration, focusing on the scope and availability of records.[8] These two examples also speak to the interdisciplinarity of approaches in this field: Legrady works in media arts and design, while Kramer-Smyth's approach is based in information management.

The approach in this project was informed by reflections on search, currently the dominant tool in the display and navigation of digital archival records. While search is a very effective technique for delivering records in response to a specific query, it has significant limitations. As an access tool, search assumes that a user is able to provide a query; but a user who is unfamiliar with the collection's scope, contents, or structure, may not be in a position to query it effectively. Personal experience suggests that such users (who are certainly in the majority) take a trial-and-error approach to search, using successive queries as a way to develop some sense of scope and context. This might be likened to using small, localised core samples to discover hidden geological features; except that in geology core

samples are used because accessing those underground structures directly is difficult and expensive. Data is, by comparison, easy and cheap to access. Visualisation enables us to literally show everything, to display large volumes of data in a way that reveals patterns and communicates context, but also provides access to the fine grain of individual elements. The work of visualisation studio Stamen Design, who make "show everything" their motto, is influential here.[9]

The central aims of this project were straightforward: to develop and trial new techniques for the visualisation of large archival collections; specifically, to create visualisations that complement (rather than compete with) search, focusing on providing large-scale, "show everything" views of the collection, and on revealing the relations and structures within it. Joanna Sassoon has argued that the push for the digitisation of cultural collections, and its focus on "content," risks a decontextualised or superficial view of the archival record.[10] A key hypothesis here was that visualisation can redress this tendency, and play a role in enhancing a sense of context in the digital collection

The research process here was practical, experimental, and iterative. A large set of "sketch" visualisations were produced, each informing the next and developing in complexity. This process was documented in detail on the project blog, which also provided a useful platform to gather feedback from peers and collaborators.[11] The final outcomes of the project – also available through the blog - were two prototype visualisations: one displaying the whole Archives collection at Series level; the other focusing on the contents of a single Series - A1.

Visualising the Collection

The first phase of the project focused on visualising the Archives' collection at the largest scale, working with data describing some 57,500 Series. This dataset reflects the Commonwealth Records System, providing a highly structured and descriptive representation of each Series, as well as recording relationships between Series, and between Series and the Agencies that control and record to them. Visualisation of this data was a stepwise process of experiment and exploration. Due partly to the sheer scale of the dataset – millions of lines of text - visualisation itself provided the most effective way to develop an understanding of the data. With each step in the process, new aspects of the dataset

became apparent, which in turn informed the next iteration of the visualisations. A short tour of these developmental stages shows how this process unfolded, as well as illustrating key features of the data and the potential role of visualisation.

The first visualisation made was a simple graph exploring the size and historical distribution of the collection (Figure 1). It is a histogram, showing the number of Series with contents commencing in each year since 1800. The dominant feature is an overall increase in the number of Series commencing per year through the course of the twentieth century. The sharp fall-off in Series after the mid-1970s can be explained by the "30 year rule", under which most recent records would be unavailable. Also notable is the small but significant distribution of Series with contents commencing before 1900. Overall, the numbers involved show the magnitude of the collection, with many years registering over one thousand commencing Series. However in its detail, this graph begins to suggest that visualisation may be able to reveal more than the broad features of the collection. Three of the sharpest spikes in the graph - years with a dramatic increase in Series contents commencing - occur in 1901, 1914, and 1939. It would seem that major historical events underlying these records can be reflected in the data; and that even the most basic visualisation can reveal traces of these events.

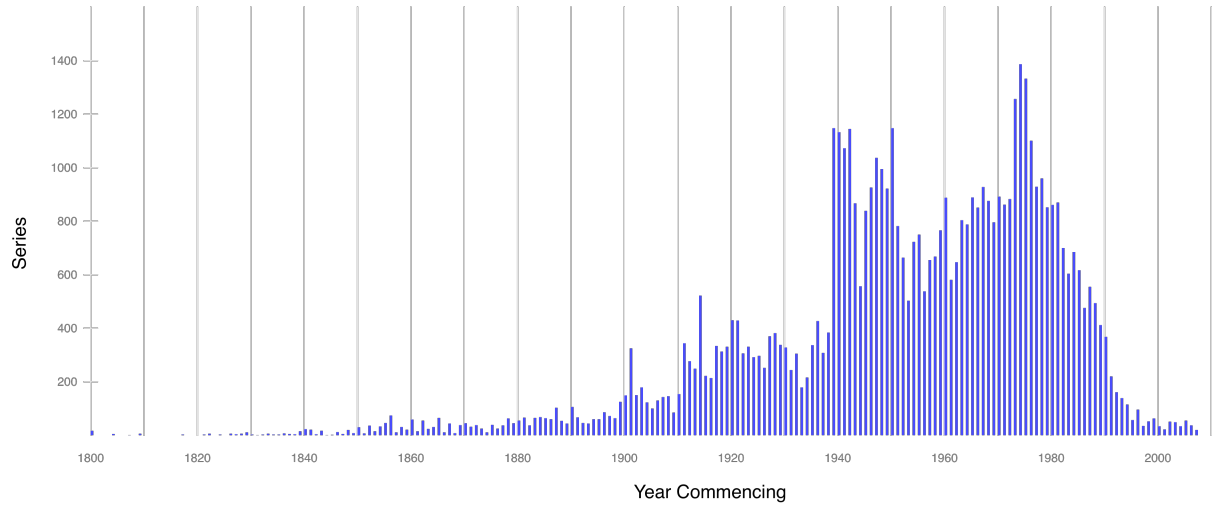

Figure 1: Histogram showing number of Series commencing per year

Subsequent experiments with the histogram form all faced a single basic problem: scale. The aim here was to create an interactive visualisation of the entire collection, at Series level; but with over 57,000 Series, the pragmatics of the "show everything" approach posed a significant challenge. The following set of sketches respond to this simply by allocating each Series its own small piece of the visual territory.

In the following sketches each Series is represented as a small coloured square on a grid (Figure 2). Series are sorted by contents start date, and layed out left-to-right, top-to-bottom. The spacing of the date labels on the vertical axis reflects the distribution shown in the first histogram, with roughly as many Series commencing between 1800 and 1900, as commenced between 1910 and 1920, and very large numbers of Series commencing in the post-War period. In Figure 2(a) colour - or more specifically hue - represents year span, with short-span Series coloured red, and long-span Series coloured purple. What results is another revealing view on the whole collection, showing for example a very large number of short-span Series, and relatively few with long spans; we can also observe features including a rapid increase in short-span Series after 1940.

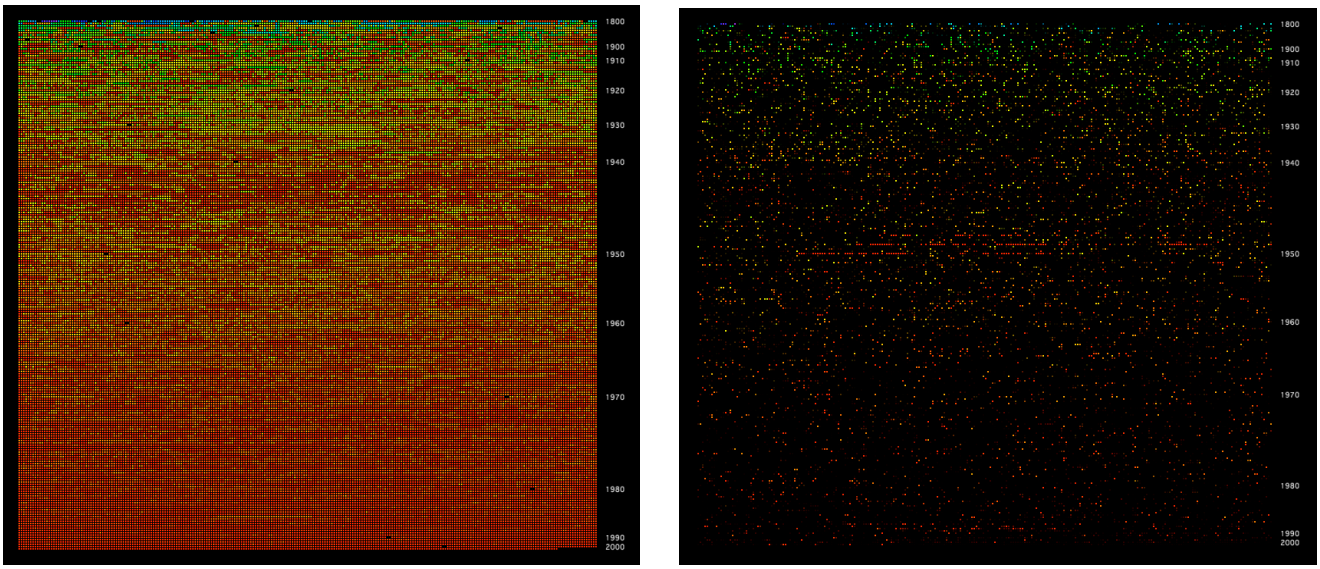

Figure 2 Grid-based visualisations showing (a) Series spans and (b) items per Series

In Figure 2(b) brightness is linked to the number of registered items in a Series; Series with few items are dim, those with many items are bright. This starry sky immediately reveals that the vast majority of Series in the collection have relatively few registered items. Its most striking features are the bands of bright red Series - short-span collections with many registered items - around 1950. Simple interaction was introduced here to browse Series details; browsing these bands reveals a group of Series documenting arrivals under the post-War Displaced Persons Program. These records stand out here not only because they contain many items, but because they are chronologically sequential, forming visual groups in the grid. This again shows that visualisation can reveal real structures within an archival collection. The addition of interaction also begins to show how a fine-grained, dynamic approach can support exploration and discovery, enabling us to quickly investigate and verify structures in the visualisation.

The key limitation of these grids is their inherent spatial constraint; despite the wide diversity of Series in the collection, each is represented here with a single, spatially uniform element. The next challenge was to use size more effectively and represent like with like - that is, link the size of a Series, to its size in the visualisation - while working within the constraints of a single-screen, whole-collection display. The solution is a spatial optimisation process, otherwise known as “packing”. In the final whole-collection visualisations Series are represented as squares whose area is proportional to both the

number of registered items it contains, and the shelf area it occupies (Figure 3). The chronological ordering of Series by year is maintained, as in the earlier grids, while the packing process re-orders Series commencing in the same year, packing them more efficiently into place. The result is a display where Series size and historical distribution are both immediately apparent. The inner square of each Series represents its number of registered items, while the outer band corresponds to the shelf space it occupies. These measures alone are revealing; for example many Series with small shelf areas but high item counts, contain items that are physically small (such as photographs or index cards).

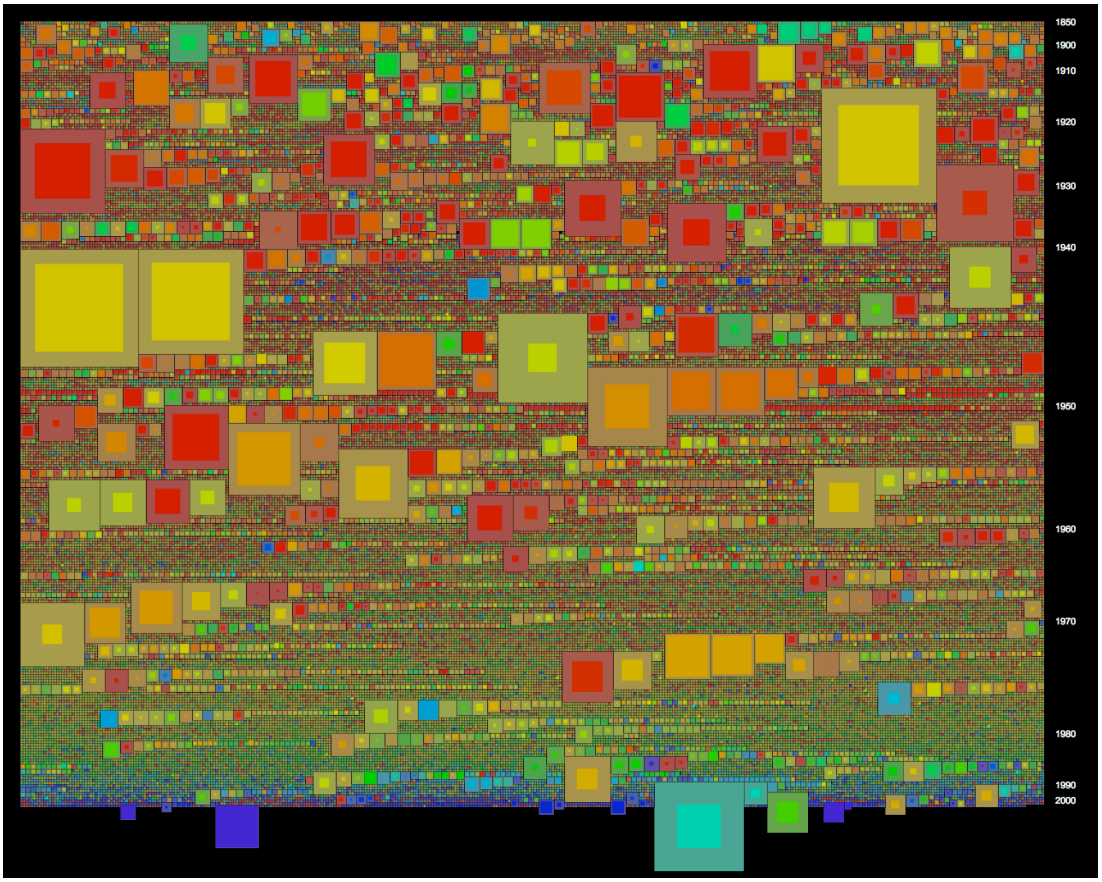

Figure 3: Packed-square visualisation

The Commonwealth Records System represents relationships between Series, and between Series and the Agencies that record to and control them. In these visualisations Agencies are shown initially as colour; the hue of each square is linked to the first listed recording Agency. Interaction enables us to add further layers of contextual data to the display (Figure 4). The user can select a Series to focus on, revealing a detailed caption with Series attributes, and a set of lines showing relationships with other

Series. These lines are colour-coded according to the CRS link types: related Series links are yellow; succession links, referring to either preceding or successive Series, are blue; controlled by links are purple, leading from one Series to another that controls (or indexes) it; and controlling links, leading in the opposite direction, are red. Browsing these links, a user rapidly gains a sense of the relationships between Series and thus the context in which a given Series sits. The recording Agencies listed in the Series caption also contribute to this sense of context; selecting an Agency highlights all its recorded Series, providing further prompts for exploration and discovery. While this final sketch has several limitations - its user interface is basic, and it lacks any ability to refine or zoom the display – it demonstrates how interactive visualisation can provide context, and enable exploration, within a very large archival collection.

Figure 4: Interactive display of Series links and Agencies

Visualising A1 – From Overview to Document

The second phase of the project moved from the whole-collection view to visualisations of a single Series, working with data from Series A1. A1 contains some 64,000 registered items, dating largely from the period 1903-1939; it was recorded to by Agencies including the Department of Home Affairs, the Department of the Interior, and the Department of External Affairs. In the dataset used here, drawn from the Archives' CRS records, each Item has a title, contents start and end dates, a control symbol, and a barcode. Other than dates, the title is most revealing of item content; this raises some interesting problems, as the title field contains unstructured text. Titles range from "August ZALEWSKI – naturalisation" to "International conference re Bills of Exchange [0.5cm]" and "Northern Territory. Pastoral Permit No.256 in the name of C.J. Scrutton."

As in the whole-collection visualisations, the initial aim here was to generate an overview of the contents of A1. The initial approach was to use simple word-frequency techniques to gain a sense of the range and distribution of text in the titles. If we take all 64,000 titles and split them into their constituent words (exluding words such as "of", "and", "to"), we can list the most frequently occurring terms, and the items that they refer to. Figure 5 shows a "word cloud" of the 150 most frequently occurring words in the list; words are sorted alphabetically, with text size linked to frequency. It is immediately clear that "naturalisation" and "certificate" occur most frequently, ahead of a wide spread of other terms. Moreover, this simple representation provides both broad coverage and a relatively fine grain. The most frequent term here, "naturalisation", occurs in some 47,000 items; while terms such as "gold" occur in only 150 items. Collectively this top 150 words refer to some 94% of the items in A1.

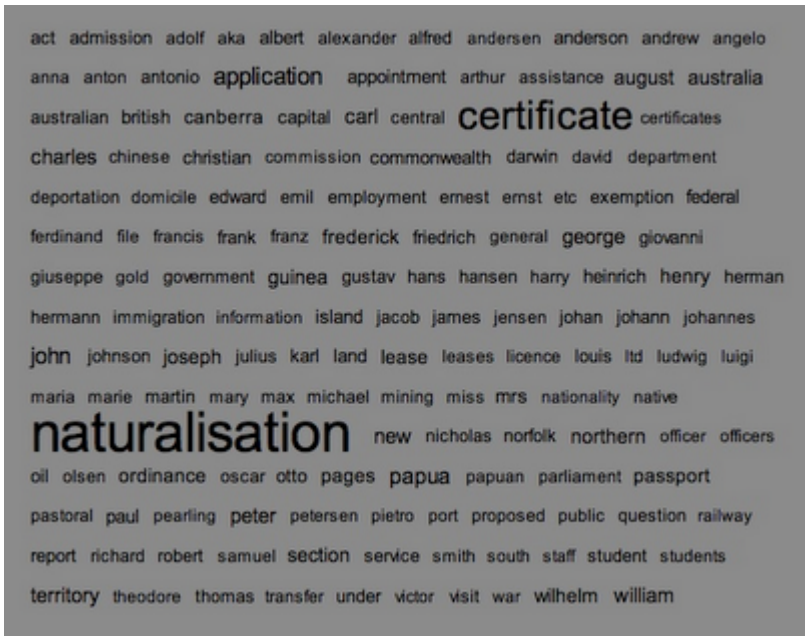

Figure 5: Word-frequency visualisation of item titles in Series A1

The following visualisations build on this simple device. Interaction offers a way to extend this static representation into a dynamic, general-purpose interface. If we add the ability to focus on or exclude terms - where focus means include *only* items containing a given term, and exclusion means include only terms *not containing* that term - we can rebuild the word cloud at each stage. This allows the user to "zoom in" on terms of interest, refining the set of items being visualised, and revealing new features within the collection. This navigation technique is simple but powerful; Figure 6 shows how the focus has been narrowed from 64000 items to less than 400, with a single click (on the term "darwin"). The rebuilt text cloud reveals new detail, terms and relationships not represented in the initial top 150 terms.

One of the risks of the word-cloud approach is that it decontextualises the source content, literally atomising it into disconnected terms.[12] In order to redress this, we can visualise the relationships between terms in a way that adds contextual information. Co-occurences are especially useful, showing which terms occur together in item titles; these links reveal, for example, that "naturalisation" and "certificate" occur together very often - not a surprise, for those familiar with the contents of A1. To add another dimension, a simple time-based histogram, shows the distribution of items over the forty-year span of the Series. Again, interaction enables exploration and discovery: hovering over a term in

the cloud highlights its distribution relative to the cloud as a whole. Finally, we can show the full details of a specific set of items, based on either title term or date. Figure 6 shows how all these features combine to support exploration and discovery. In this case we have focused on the term "darwin" to discover a dramatic spike in the histogram - a large increase in the number of items occuring in 1937. Hovering over terms in the cloud offers some clues; we can see the strong co-occurrences of "darwin", "cyclone", "march", and "1937". Finally the listing of item titles confirms our developing hypothesis; a cyclone did hit Darwin in March 1937, as these records show.

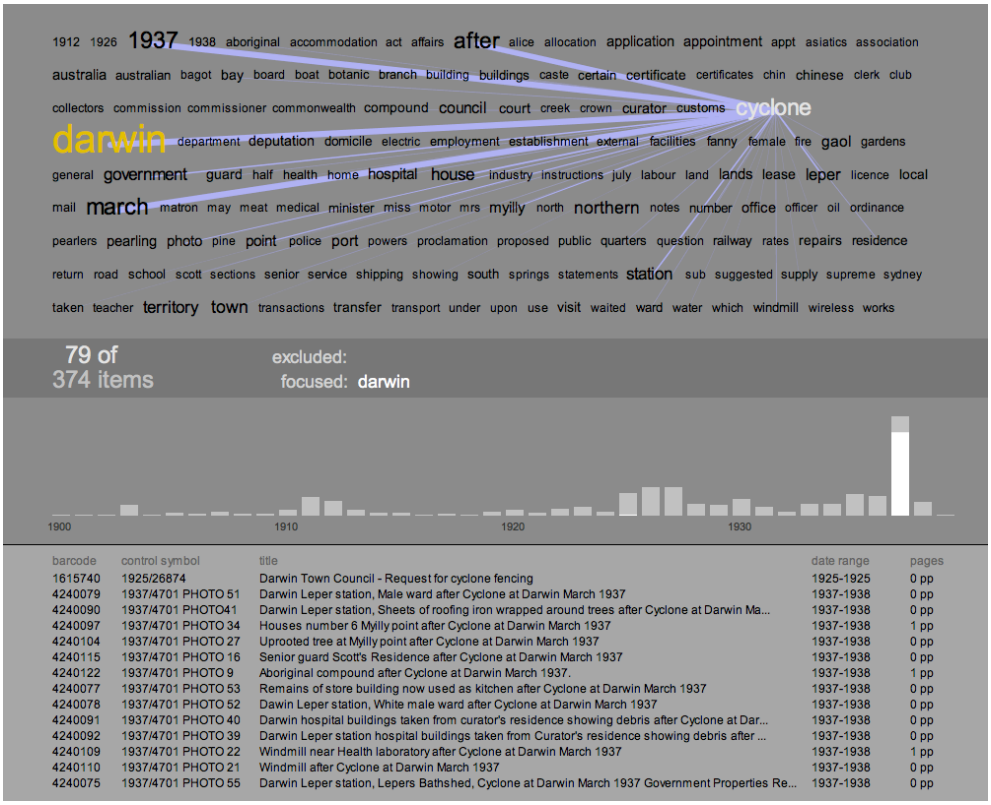

Figure 6: A1 Explorer interface, showing word cloud, co-occurrences, year histogram, and item listing

The final challenge in this process was to zoom in again, to the level of the individual document. The National Archives has digitised a significant portion of its records: it currently stores 18.2 million images, accessable through its RecordSearch service, including many of the items in A1. This prototype loads page images from the Archives servers over a network connection, enabling a user to

move rapidly from a synoptic overview of the collection to a close investigation of a specific document. For example, we can readily move from finding the abstract data-traces of the 1937 Darwin cyclone, to viewing photographs of the storm damage. As Figure 7 shows, the materiality of the documents can be striking – in this image we see a page from item 1921/22488 – "Pearling Lugger Stolen by Japanese Thursday Island" – showing the handprint of one Unoske Shimomura. Such documents are the core of the collection; what this visualisation shows is how computational techniques can support and richly inform our navigation of this collection, as well as, crucially, providing access to its primary materials.

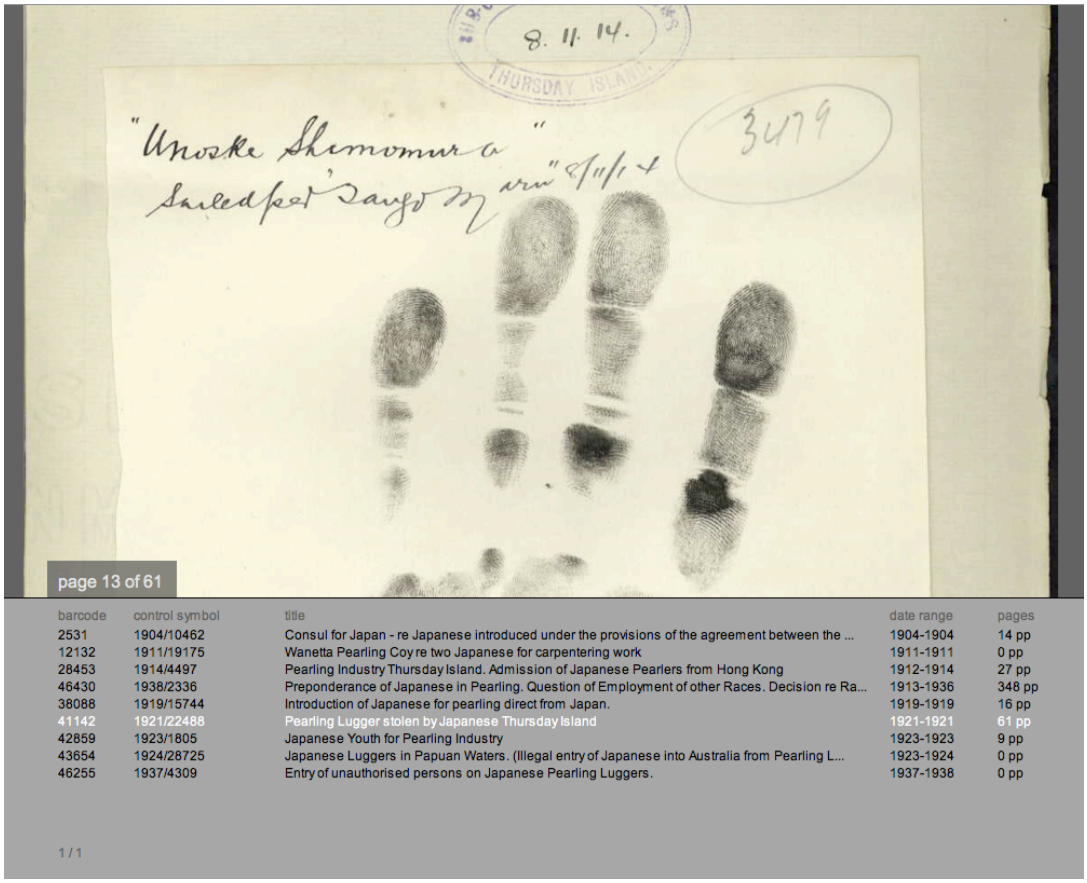

Figure 7: A1 Explorer interface, showing document image display.

Reflections and Implications

The aim of this project was to develop visualisation techniques to reveal structure and context within archival datasets, in a way that supports research, exploration and discovery, and provides a wide view to complement the targeted focus of search. The key outcomes of the project - two interactive prototypes - demonstrate a range of techniques that achieve this aim, at two different scales within the collection. Collectively, they show the potential for interactive visualisation to support exploration at every level from the entire collection - millions of items - to a single item.

The central challenge in these visualisations was to make compact but effective representations of very large collections. The final prototypes show two different solutions to this challenge: the Series browser follows Stamen Design's "show everything" approach, displaying the entire collection in a single static image, with an interactive overlay of additional context. This approach provides a stable, unified representation of the collection, but the sheer number of elements it presents remains a challenge for the user. On the other hand the Items browser creates a compact text-cloud synopsis that is far easier to navigate, but cannot provide a complete representation of the Series content. The Items browser also shows how effective simple interaction can be in refining the scope of exploration; future work will focus on applying this approach to the whole-collection view to create a more navigable interface.

The implications of this project extend beyond the design of the visualisations themselves. In line with the context outlined earlier, this project demonstrates the increasing accessibility of visualisation tools and processses. Public access to public sector data seems also likely to improve, given initiatives such as Data.gov in the United States, and the Government 2.0 taskforce in Australia.[13, 14] In this context - increasingly available datasets with broad public significance, and increasingly accessable tools for data visualisation - we can expect to see many more visualisations of digital cultural collections, and other public datasets; interactive visualisation will play a crucial role in facilitating access to and exploration of this data. For archival collections, this project also hints at the potential of visualisation as a research tool for the users of collections, in particular its ability to reveal pattern and structure within archival data. The spikes in the very first histogram, or the prominence of the Darwin cyclone, suggest that in some cases historical events can become legible through the visualisation of their

records. As Martyn Jessop argues, visualisation in the humanities is "a scholarly activity rich in opportunities to create knowledge", though these new methodologies also demand careful analysis, and new computational and visual literacies.[15]

It seems likely that our archives and cultural collections will be increasingly, if not exclusively, digital collections. If so, the question of how those collections are displayed, represented, and accessed, is a crucial one. Current techniques, such as search, are effective in some instances, but limited in others. This paper has demonstrated one of the great opportunities of the digital collection, showing how the computability of data, and its polymorphous plasticity, can support new techniques for representing archival collections. The prototypes presented here are in no sense definitive; they are early experiments in what should be an ongoing exploration, for the stakes are siginificant. If the records of our society take the form of an immense field of data, our ability to represent, navigate and explore that field will be crucial to our understanding of not only the data, but of society and culture itself.

---

1. Abby Smith, *Why digitize?* Council on Library and Information Resources, Washington, D.C., 1999.

2. Michael Friendly and Daniel J Denis, "Milestones in the history of thematic cartography, statistical graphics, and data visualization." (2006) at http://euclid.psych.yorku.ca/SCS/Gallery/milestone/milestone.pdf.

3. Vitaly Friedman, "Data Visualization: Modern Approaches," *Smashing Magazine*, 2 August 2007, at http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches/

4. See for example Daniel Cohen et al, "Interchange: The Promise of Digital History," *The Journal of American History* vol. 95, no. 2, September 2008, at http://www.journalofamericanhistory.org/issues/952/interchange/index.html.

5. Martin Jessop, "Digital visualization as a scholarly activity," *Literary & Linguistic Computing* vol. 23, no. 3 (September 2008): 281-293.

6. Lev Manovich, "Visualizing Cultural Patterns," *databeautiful*, 23 May 2008, at http://databeautiful.net/2008/05/23/visualizing-cultural-patterns/.

7. George Legrady, "Making Visible the Invisible," 2005, at http://www.mat.ucsb.edu/~g.legrady/glWeb/Projects/spl/spl.html.

8. Jeanne Kramer-Smyth, "ArchivesZ: Visualizing Archival Collections," 2007, at http://archivesz.com/.

9. Matt Jones, "Data as Seductive Material," March 2009, at http://www.slideshare.net/blackbeltjones/data-as-seductive-material-spring-summit-ume-march09.

10. Joanna Sassoon, «Documenting Communities: If digitisation is the answer, what on earth is the question?», *Connections and Conversations*, Australian Society of Archivists conference, Port Macquarie, 2006, at http://www.archivists.org.au/files/Conference_Papers/2006/Sassoon_ASAConference2006.pdf

11. Mitchell Whitelaw, *The Visible Archive*, 2009, at http://visiblearchive.blogspot.com

12. Jodi Dean, "Tag clouds and the decline of symbolic efficiency," *I cite*, January 22, 2009, at http://jdeanicite.typepad.com/i_cite/2009/01/tag-clouds.html.

13. *Data.gov,* 2009, at http://www.data.gov

14. Government 2.0 Taskforce, 2009, at http://gov2.net.au

15. Jessop, op. cit.