Exploring the digital commons: an approach to the visualisation of large heritage datasets

Sam Hinton
Media Arts & Design, University of Canberra
Canberra 2601, Australia
Sam.Hinton@canberra.edu.au

Mitchell Whitelaw
Media Arts & Design, University of Canberra
Canberra 2601, Australia
Mitchell.Whitelaw@canberra.edu.au

Visualisation of complex datasets is often designed to assist communication and to make that data more visually accessible (Friendly and Denis, 2006). In some recent approaches to data visualisation, the goal of visualising datasets is not to reveal a single underlying 'truth' that hides in complex data, but rather to visualise the structure of the data itself, to 'show everything' and see what emerges (Jones, 2009).

The latter approach is particularly useful in the visualisation of large digital heritage collections, which present challenges for conventional data visualisation because they are often polymorphous and idiosyncratic. Interactive tools for exploring heritage datasets can enable people to explore and play with potential relationships between parts of the collection and to learn about the collection itself and thus better understand the material it contains and how that material has been organised.

This paper provides a tangible demonstration of this approach and how it has been embraced in two recent interactive heritage collection visualisation projects: Whitelaw's Visible Archive (which visualises the collection of the National Archives of Australia) and Hinton and Whitelaw's Flickr Commons Explorer (which visualises nearly 40 photographic collections comprising more than 20,000 images available through Flickr).

Digital collections. Data visualisation. Flickr. Archives. Java. Processing

1. DATA VIZ AND THE SHOW EVERYTHING APPROACH

As Friendly and Denis show, the visual display of quantitative information has a long history, spanning cartography and scientific diagrams, graphs and charts. (Friendly & Denis, 2006) 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. (Friedman, 2007) Current visualisation practice is a broad and growing field spanning information technology, the digital humanities, design, and art.

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. (see, for example, Cohen, 2008; Jessop, 2008). Researchers in the field have begun to recognise the potential of visualisation; Lev Manovich (2008) for example describes research into 'visualising cultural patterns'. Examples of visualisations of non-scientific data abound, some well-known examples including Stamen Design's *In The News*, Harris and Kamvar's *We Feel Fine* and Borevitz's *State of the Union*. The latter of these provides a fascinating visualisation of every state of the union address given by US presidents, providing a compelling insight into the way that political agendas change from year to year, from administration to administration and across the decades. Borevitz's work is more than a presentation – it's a tool that allows the viewer to explore the state of the union addresses, encouraging them to develop a sense of the entire corpus, but also beckoning them deeper, perhaps to engage with the original texts themselves.

Examples of visualisations of large heritage 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. Jeanne Kramer-Smyth's ArchivesZ project (2007) is more relevant to the projects discussed in this paper – an interactive tool using visualisation to support search and exploration, focusing on the scope and availability of records. 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 the projects below 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 (Jones, 2009).

2. DIGITAL HERITAGE COLLECTIONS AS DATA SOURCES

As the costs of computing and data storage have fallen, museums and other cultural institutions have embarked on a process of collection digitisation. Across the world millions of photographs and manuscripts are being scanned, their digital copies stored in databases where they can be accessed almost instantly with the correct keywords. As more and more material is digitised, questions about how best to use these digital materials, and how best to make them accessible, are becoming more pointed. At the same time cultural institutions are becoming more concerned with engaging their visitors, giving people the tools to work with collections to construct their own pathways and develop their own perspectives on the material rather than providing them with a single institutionally constructed view. (Kelly, 2006)

While digitisation of collections is important for reasons beyond accessibility (conservation, for example), the value of being able to share a large collection with millions of people across the internet, and the capacity this gives cultural institutions to engage with the public, is difficult to overstate. As Aljas and Pruulmann-Vengerfeldt (2009, p. 61) point out, however, there have been numerous studies that criticise cultural institutions for developing collection databases that are isolated from a consideration of its users. Digital materials only become useful if people can access them in a meaningful way; simply making digital copies of materials available through a search box located on a web interface is a start but, as noted above, is problematic.

Many cultural institutions appear to be rising to the challenge and are looking toward the web and social media technologies like Facebook, Twitter and Flickr as a means of extending their collections into the places 'where [people] already work and play' (Kalfatovic et al. 2008). The image of a 21st century museum that maintains a physical location but also reaches out into the world through the internet is beginning to take shape. But beyond extending the reach of museums though the internet, social networking emphasises user generation of content. People no longer simply view or consume cultural content; they make it, re-use it, and annotate it, adding meaning and creating new derivative media forms. This emerging mode of use is addressed, for example, in the New Literacies New Audiences project which explores what social media means for cultural institutions, and how visitors may build different and personalised ways of accessing or organising their collections. (Russo & Watkins, 2008)

The Flickr Commons project provides an excellent and highly relevant example of how institutions may make use of social media. Initially a joint effort between the U.S. Library of Congress and Flickr, the Flickr commons brings together photographic collections from more than 30 cultural institutions around the world. The project has been well documented, culminating in a significant report from the Library of Congress, as well as research papers from participating organisations such as The Smithsonian (Kalfatovic et al., 2008). The Commons now boasts tens of thousands of photographs, all presented and accessible through the Flickr interface. Flickr provides the search and browsing functions that allow people to find photos in the collections, but more importantly it allows people to engage with the content and enhance it. For example, people can leave comments about

photos, annotate regions within an image, and create folksonomies (user generated keywords) that form crowd-sourced descriptions of individual photos. The value of this latter approach is described on Sydney's Powerhouse Museum web site: 'Sometimes museums describe objects in language that is highly specialist and user added keywords are useful in bridging the 'semantic gap' between the language of the museum and that of the user' (Powerhouse, 2010).

Perhaps most importantly in the context of this paper, the Flickr interface can be accessed via an API (application programming interface) – a kind of gateway that allows third party applications to access the Flickr collection programmatically. When heritage collections are available in such an open, structured and well documented manner, they become rich cultural datasets. For the visualisation designer, heritage collections represent an intriguing combination of opportunity and challenge – challenge because of their unevenness (compared with scientific data sets) and opportunity because the content comprises the material traces of society and culture itself; their diversity and complexity reflects the complexity of the social, cultural, historical and institutional systems that created them.

3. CASE STUDIES: VISIBLE ARCHIVE AND COMMONSEXPLORER

The following case studies illustrate the practical application of the concepts described above. The development of two distinct but related projects is presented – the A1 Explorer from Whitelaw's Visible Archive project, and the commonsExplorer (http://creative.canberra.edu.au/cex/).

The projects were developed using the Java programming language and Fry and Reas' Processing extensions, plus some other open source libraries, mainly because of ease of use and the ready availability of software libraries for accessing data sources programmatically. As an environment for data visualisation, Processing provides an invaluable set of tools for graphics and data processing, and supports an active and supportive community that blends art and design with computer science.

Both projects share the common design challenge of presenting a large archive of digital material in a way that encourages exploration of the collections. The guiding principle was to show everything, make few assumptions about how users will use the materials, and to provide an interface that is visually appealing but which is as unobtrusive as possible.

The goal in these projects is not only to visualise data structures in heritage collections, but to engage the visualisation as a tool for data exploration. Our aims follow what Keim (2001) describes as 'visual data exploration' – a visualisation approach that engages the user in an exploration of large datasets, providing insights and encouraging the user to form and test their own hypotheses. Our practical approach aligns with Scheiderman's influential 'visual information seeking mantra': 'overview first, zoom and filter, then details-on-demand' (Schneiderman, 1996, 337). The projects utilise both structured information (such as unique document identifiers) but importantly, also use unstructured data (such as words from document titles, or image data from thumbnails) as 'clues' that allow the user to develop hypotheses about the data and thus to discover their own pathways into the collections. The tools also embrace the idea that while exploration is rewarding, access to the source document is ultimately the most satisfying experience, and so both projects provide the user with the capacity to move from broad overview to source document in a single environment.

3.1 The Visible Archive

Supported by the National Archive of Australia, the Visible Archive was a practice-led research project in the visualisation of heritage collections. The aims of the project were to create prototype visualisations of the Archives collection at two scales: the whole collection – comprising some 60,000 archival series – and a single series, containing around 64,000 documents (Whitelaw, 2009b). The discussion here will focus on the single series visualisation, illustrating our interest in zoomable representations that move easily from whole collection to single document.

The National Archive's Series A1 contains some 64,000 registered items, dating largely from the period 1903–1939; it contains records from Australian Federal agencies including the Department of Home Affairs, the Department of the Interior, and the Department of External Affairs. In the dataset used here – a subset of the fields in the Archives' own records – each item in the set has a title, contents start and end dates, a control symbol, and a barcode. Other than the (well-structured) 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'.

The initial challenge here was to generate an overview of the contents of series A1. Our

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 (excluding uninformative (or 'stop') words such as 'of', 'and', 'to'), we can list the most frequently occurring terms, and the items that they refer to. Figure 1 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. Notably, this compact representation provides both broad coverage of the series contents, 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 these 150 words refer to some 94 per cent of the items in A1. This efficiency is related to a statistical property of natural language known as Zipf's law, which observes that the prevalence of a term is inversely proportional to its rank on the frequency table – so the most frequent term occurs around twice as often as the second most frequent, and so on (Zipf, 1949).

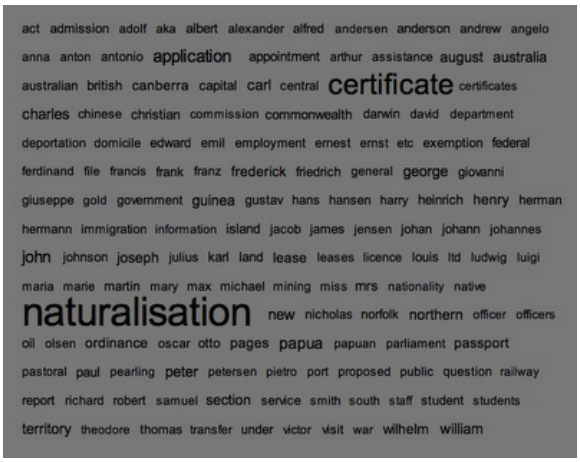

Figure 1: *Word-frequency visualisation of item titles in Series A1*

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 term not containing that term – we can rebuild the word cloud interactively. 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 2 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. (Dean, 2009) In order to redress this, we can visualise the relationships between terms in a way that adds contextual information. Co-occurrences 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 2 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 occurring in 1937. Hovering over terms in the cloud offers some clues; we can see the strong 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.

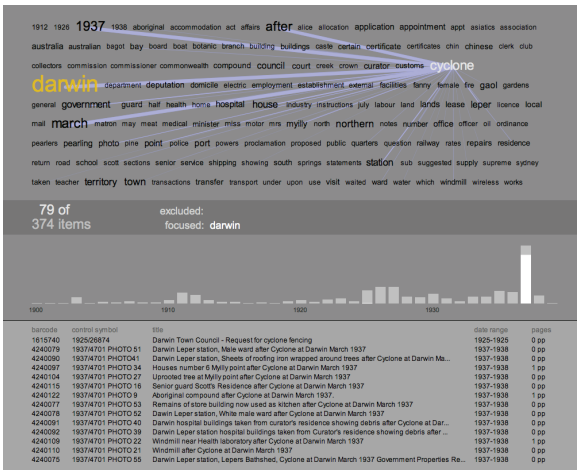

Figure 2: *A1 Explorer interface, showing word cloud, occurrences, year histogram, and item listing*

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, accessible 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 3 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.

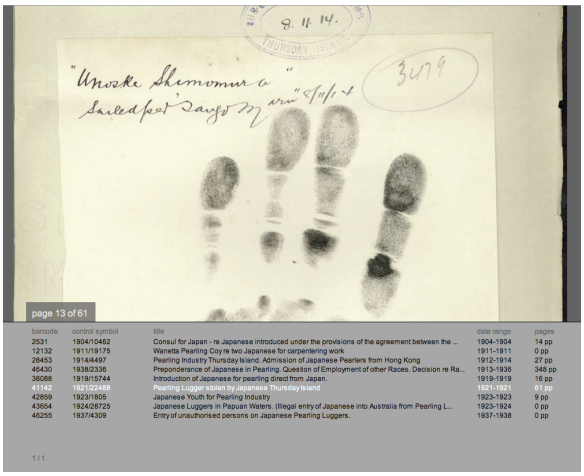

Figure 3: *A1 Explorer interface, showing document image display*

Finally, the Visible Archive project also hints at a more expansive possibility for the visual exploration of heritage datasets. These datasets are formed by complex social, cultural, historical and institutional systems and forces – if the data bears traces of those wider forces, can visualisation help reveal them? The very first visualisation made in this project was a simple graph exploring the size and historical distribution of the collection (Figure 4). 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 – not surprising given the growth in Australian population and government during that time. However in its detail, this graph also seems to show traces of those wider forces. Three of the sharpest spikes in the graph – years with a dramatic increase in Series contents commencing – occur in 1901, 1914, and 1939. 1901 was the year of Australian Federation; 1914 and 1939 mark the beginning of two World Wars. 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.

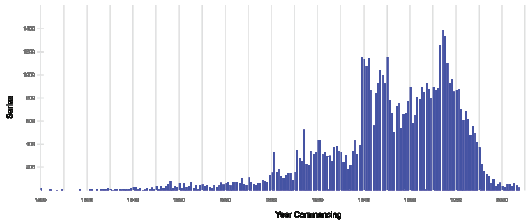

Figure 4: *Histogram showing number of Series commencing per year*

3.2 Flickr Commons Explorer

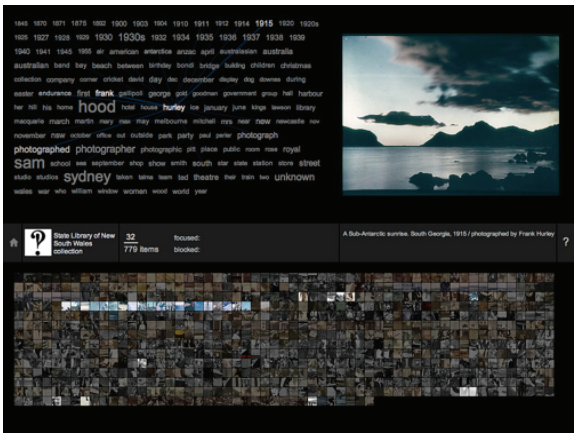

Figure 5: *Flickr Commons Explorer's main interface*

Building on techniques developed in the Visible Archive project, the commons Explorer was initially developed in response to Mashup Australia, a competition run by the Federal Government's Gov2.0 Taskforce. The competition sought to demonstrate the value of making government data sets accessible to the public and to showcase the range of creative applications of public data made possible by simply 'putting it out there'. Among the data sources available from a variety of government agencies was material from a number of Australian cultural institutions that had become part of the Flickr Commons. Our initial focus was the Flickr set of the State Library of New South Wales (NSW), and the aim to produce rich, dynamic, ‘show everything’ interface, applying techniques from the Visible Archive work to image-based content. As the project developed however, the scope broadened to include the whole Flickr Commons - at the time of writing 34 collections and over 40,000 images.

The Explorer presents a three-pane interface consisting of a term cloud, a single image view and a thumbnail grid, with a central strip providing navigation and orientation (Figure 5).

Figure 6: *term cloud*

As in the A1 visualisation, the term cloud shows the 150 most frequently occurring words in the titles of the current set of images, with co-occurrence lines, and a similar navigation mechanism based on focusing and blocking title terms. Commons Explorer shows that this technique can be effectively applied across a range of different heritage collections. Word co-occurrences are a powerful cue for hypothesis-forming, and the descriptive titles of these items provide rich material for this approach.
It's important to note here that the term cloud is based solely on image titles, rather than user-generated tags or other photo descriptors. There are a number of functional reasons for this decision, the chief one being that other photo descriptors available have patchy coverage. For example while every Commons image has a title, not all have tags. A possible future enhancement is to mingle user-defined keywords with title terms, allowing both sources to describe photographic records.

Here again the word-frequency cloud provides a compact overview of large collections. Once again, according to Zipf's law, the top-level cloud of 150 words almost always refers to more than 75 per cent of the images in the set - even in a collection numbering in the thousands. Interestingly however there are significant differences in coverage between collections, reflecting diversity in both titling and collection structures. Most collections contain significant clusters of similarly-titled items, while a few are more heterogeneous. For example the top-level cloud of the State Library of NSW collection refers to 97 per cent of its 810 items; while for the Queensland State Library it refers to 90 per cent of much a smaller set of 543 items. Meanwhile for the much larger US National Archives collection the top 150 terms refer to 94 per cent of its 4,725 items. At the other end of the scale the top-level cloud covers only 53 per cent of the mere 213 items in the National Archives UK collection - this seems to be a product of both a very diverse collection, and relatively terse titles. Although it seems to be an unusual case, this raises an important question which has yet to be addressed – of how to communicate the notion of coverage to the user, and make those items not reflected in the top level cloud more immediately discoverable. After all, the outliers or exceptional items in a collection may well be among the most interesting.

The thumbnail grid element of the Explorer is an attempt at a 'show everything' image visualisation that can scale from tens to thousands of elements. As the number of elements grows, the grid size decreases to fit in the available space, but rather than scale images down, we simply crop the thumbnails – the intention isn't to represent the whole image, but to provide rich but unstructured visual clues: a visual core sample through the whole set. The results show how this can help reveal structure within the collection. Different photographic processes are instantly apparent – monochrome, sepia, cyanotype, stereoscopic, Kodachrome. Other similarities are also apparent, even in small tiles - it's often possible to distinguish landscapes, from portraits, from images of documents, for example. Groups of images with similar subject matter jump out; a striking example is the luminous Antarctic blue of Frank Hurley's photographs in the State Library of NSW collection (Figure 5). Dates form a related layer of structure in the grid: many collections, including the initial State Library of New South Wales set, include dates in image titles. We automatically find and parse these dates, and sort dated images first in the grid. This approach is simple, and prone to occasional false positives, but it degrades gracefully – any items without dates are presented as ordered in the Flickr collection.

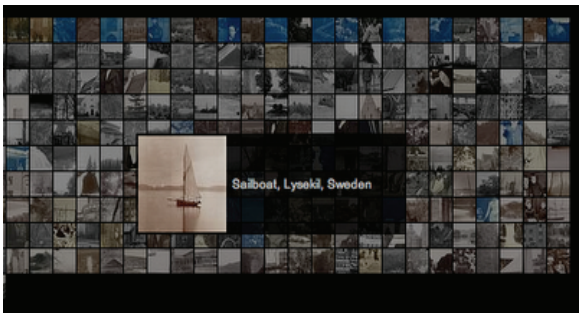

Figure 7: *thumbnail grid*

Like the word cloud, the grid provides what Keim (2001) calls 'details on demand', revealing the full thumbnail and image title on mousing over a tile. This contributes to the process of hypothesis-forming and insight; by showing the part-whole relationship between one tile and its full thumbnail, the user is better able to hypothesise about the significance of other tiles. Thus the application presents the user with a rich mass of partial

information - or rather data: linked fragments of titles, and of images. Moments of discovery come when we see those fragments unified in a source image: the fragments are contextualised and become more meaningful. This contextual information then propagates back to the fragmentary display – when it works best there is a feedback loop from hypothesis forming, to discovery, to context and back to hypothesis forming. Whitelaw (2009a) has argued elsewhere for a conceptual distinction between data and information, which is relevant here: these fragments are data points, abstracted and decontextualised. Information occurs only when we link and interpret those fragments – and it happens strictly on the human side of the screen. Even before we apply it as a methodology, Schneiderman's (1996) 'visual information seeking mantra' reminds us that information is sought, rather than provided in advance. This is an important feature of the Explorer - it emphasises user interpretation over programmatic or institutional interpretation, and aims to engage the user in a virtuous cycle of curiosity, hypothesis-forming and discovery. In doing this the Commons Explorer has achieved what we set out to do. It provides a rich experience that encourages an understanding of context, and enables discovery in large collections. We've also shown that this approach is broadly applicable - it could be applied usefully to other large image collections as a browsing tool, including potentially collections stored in other open network accessible locations.

4. CONCLUSIONS

As cultural institutions make their collections more readily available in a digital form, there is a growing acceptance that simply putting data online is not enough; there needs to be ways that people can engage and be creative with the digital material. Data visualisation techniques, like those described in this paper offer one method of providing information to people in a way that allows them to understand and engage with digital collections, to play with them and to develop a sense of the collection's scope. We also believe that a show everything approach that attempts to lay bare the content and structure of a collection offers an approach that is broadly compatible with many cultural institutions efforts to engage with the public and make themselves open and accessible in an online environment.

It is worth noting that these visualisations owe a great deal to the openness of the subject organisation's collections. For Whitelaw's Archives Explorer the openness was a result of direct engagement with the institution. Likewise, the Flickr Commons project provided much more than a web interface for accessing photos – it also provided this open well documented API, which made the construction of the Flickr Commons Explorer possible. There is a great deal of value when collections are placed online in an open and flexible manner, with access through open well-documented APIs or storage systems that encourage rather than prevent access to a collection's data structures.

The Flickr Commons Explorer (and source code) is available for download (for Windows, Linux and Mac) from http://creative.canberra.edu.au/cex.

5. REFERENCES

Aljas, A. and Pruulmann-Vengerfeldt, P. (2009) Digital cultural heritage: Challenging Museums, Archives and users. In Nico Carpentier et al. (eds), *Communicative Approaches to Politics and Ethics in Europe*. Tartu University Press.

Cohen, D. et al. (2008) Interchange: The Promise of Digital History. *The Journal of American History* 95:2. http://www.journalofamericanhistory.org/issues/952/interchange/index.html (30 March 2010).

Commonwealth of Australia (2009) Mashup Australia. http://mashupaustralia.org/ (20 January, 2009).

Commonwealth of Australia (2010) Government 2.0 Taskforce. http://gov2.net.au (30 March 2010).

Dean, J. (2009) Tag clouds and the decline of symbolic efficiency. I cite. http://jdeanicite.typepad.com/i_cite/2009/01/tag-clouds.html (30 March 2010).

Friedman, V. (2007), Data Visualization: Modern Approaches. *Smashing Magazine*, 2 August 2007 http://www.smashingmagazine.com/2007/08/02/data-visualization-modern-approaches/ (30 March 2010).

Friendly, M. and Denis, D. J. (2006) Milestones in the history of thematic cartography, statistical graphics, and data visualization. http://euclid.psych.yorku.ca/SCS/Gallery/milestone/milestone.pdf (30 March 2010).

Fry, B and Reas, C. (2010) Processing. http://www.processing.org/ (30 March 2010).

Jessop, M. (2008) Digital visualization as a scholarly activity. *Literary & Linguistic Computing*, 23:3, pp. 281–293.

Jones, M. (2009), Data as Seductive Material. http://www.slideshare.net/blackbeltjones/data-as-seductive-material-spring-summit-ume-march09 (19 February 2010).

Kalfatovic, M. et al. (2008) Smithsonian Team Flickr: A Library, Archives, and Museums Collaboration in Web 2.0 Space. *Arch Sci*, 8, pp. 267–277.

Keim, D. (2001) Visual exploration of large data sets. *Communications of the ACM*, 44, pp. 38–44.

Kelly, L. (2006) Museums as Sources of Information and Learning: The Decision Making Process. *Open Museum Journal*, 8. http://hosting.collectionsaustralia.net/omj/vol8/kelly.html (19 January 2009).

Kramer-Smyth, J. (2007) ArchivesZ: Visualizing Archival Collections. http://archivesz.com/. (30 March 2010).

Legrady, G. (2005) Making Visible the Invisible. http://www.mat.ucsb.edu/~g.legrady/glWeb/Projects/spl/spl.html (30 March 2010).

Manovich, L. (2008) Visualizing Cultural Patterns,. Databeautiful. http://databeautiful.net/2008/05/23/visualizing-cultural-patterns/ (30 March 2010).

Powerhouse Museum (2010) http://www.powerhousemuseum.com/collection/database/browsekeywords.php (March 30 2010).

Russo, A. and Watkins, J. (2008) New Literacy New Audiences: Social Media And Cultural Institutions. Chartered Institute for IT. http://www.bcs.org/server.php?show=conMediaFile.8816 (5 February 2010).

Shneiderman, B. (1996). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In P*roceedings of the 1996 IEEE Symposium on Visual Language.* http://portal.acm.org/citation.cfm?id=834354 (29 March 2010).

Smith, A. (1999) Why digitize? Council on Library and Information Resources, Washington, D.C.

Whitelaw, M. (2009a) Art Against Information: Case Studies in Data Practice. *Fibreculture*, 11. http://journal.fibreculture.org/issue11/issue11_whitelaw.html (14 May 2009).

Whitelaw, M. (2009b) Visualising the Digital Archive: the Visible Archive project. *Archives & Manuscripts*, 37, pp. 22–40.

Zipf, G. K. (1949) *Human behavior and the Principle of Least Effort; An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, Mass.