AICCM Contexts for Conservation

2013 National Conference - Adelaide 23- 25 October

Data visualisation: a new tool for conservation

Peter Shaw, National Archives of Australia

Abstract

In the National Archives of Australia we are flooded with information in databases, spread sheets and images. We use this data to identify, understand and retrieve collection items. Analysis of collections through physical surveys and database interrogations is used to establish conservation priorities and undertake identified projects. This methodical approach can be time consuming and labour intensive. Web based word searches can be a little like viewing the world through a letterbox. By harnessing data sets we can use visualisation techniques to look at *everything* and reveal associations that can be used for preservation and access purposes. Visualisation can be a useful tool for access, preservation planning and data entry quality. Five years ago, the Archives commissioned a project called the Visible Archive, in which visualisation techniques were applied to large Archives datasets to develop ways of exploring and understanding the contextual relationships of records. The Archives is currently supporting a project to explore content based image retrieval that will gather like images in a data set. For conservators and archivists this will have the benefit of identifying duplicate images; associating images that may be split across collections and allowing data entry comparison for consistency, accuracy and completeness. I aim to demonstrate that with available data sets conservators can use visualisation processes to identify conservation priorities in image collections and target records at risk and known to be deteriorating.

Introduction

My aim is to propose a different approach to the conservation and management of photographic collections. I want to look at ways that we conservators can use digital tools that enhance the way we work with large collections.

We live in a world inundated by data about all sorts of things.

At the National Archives of Australia we hold over 40 million records that are housed in 378 kilometres of shelving. Over 8 million items are listed on our RecordSearch database and we have about 330,000 publicly accessible photographic images.

How does the Archives prioritise the conservation and digitisation of thousands, (or millions) of images, with limited resources? We use a number of strategies, including a mass preservation approach of developing and providing appropriate storage facilities for long-term preservation. There is a continuing program to relocate photographic materials into storage areas with environmental conditions more suitable for long-term preservation. Other standard basic conservation measures are to re-box photographic records into archival containers and repackage them into photographic activity tested (PAT) sleeves. We undertake physical collection surveys, and sometimes include vinegar syndrome testing by item or box level. There are limitations with any approach taken. Physical surveys are labour intensive and can be either, narrow and deep (item condition testing), or broad and shallow, (a survey to determine repackaging needs). Repeatability of vinegar syndrome testing at box level can be unreliable or misleading. Especially, if later re-testing cannot establish whether individual records are stable or have continued to deteriorate in cold storage.

Intellectual collection surveys are also undertaken, primarily by interrogating the Archives' RecordSearch database. These use record descriptions either provided by the originating agency or added later by Archives' staff. This information can be useful in directing and prioritising photographic preservation work, but it is not always consistent because of different data entry standards used over time. Controlled vocabularies give more consistent results, but are limited by length and consequent comprehensiveness. If a researcher is not aware of particular accepted use or technical terminology, then records will be missed in a word-based search.

Conservation of archival photographic collections is necessarily fragmentary, concentrating on those records that are found to be deteriorating or have been identified as vulnerable formats. Because of limited resources, photographic records undergo intensive conservation treatment only if they are unique or have high value as an artefact. This approach constitutes only about six per cent of the volume of photographic preservation work undertaken each year. Digitisation of photographic records is the other major intensive photographic preservation treatment undertaken at the Archives. Over the past decade about 330,000 or 10 per cent of the photographic records have been digitised. This is done to a standard that allows the original record to be safely housed in conditions suitable for long-term preservation and provides a high-quality digital copy that is expected to satisfy access requests.

With so many digital image files created by different technical staff over time it can be expected that there will be anomalies in the image collections. These include duplicate images coming into custody, image sequences disrupted and split over different records' transfers from agencies, and different or inadequate descriptive standards applied to groups of similar images. These issues can affect the usefulness of, and accessibility to, collections and undermine or devalue the preservation effort.

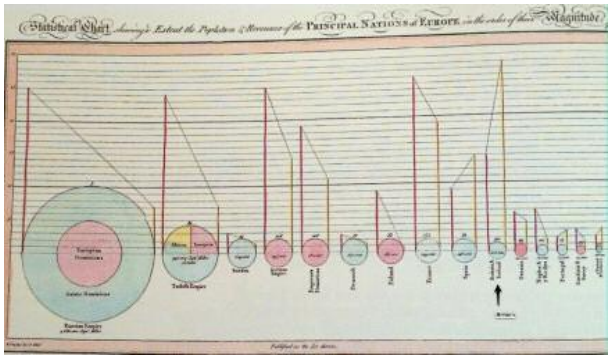
The power of visualisation

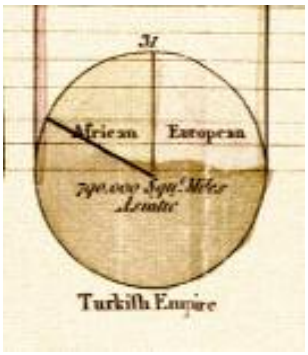
I want to discuss some of the opportunities and possibilities of using visualisation as a practical and effective tool for collection analysis. Data visualisation can give us real time, not just historical, views of data.

Visual memory is very powerful and has its roots in antiquity. The art of memory was cultivated by the Greeks and Romans. In 55BC Cicero credited the Greek poet Simonides with the invention of (what was later called) artificial memory.

Simonides invited to a banquet, recited a lyric poem in honour of his wealthy host Scopas, in which he included a passage honouring Castor and Pollux, twins of Greek mythology. Afterwards Scopas meanly offered him only half his fee and told him he could ask for the other half from Castor and Pollux. A little later, a message came to Simonides that two men were waiting outside for him. However, when he went outside there was no one there. Shortly afterwards the house in which he had been dining collapsed and the people inside were crushed beyond recognition. When their relatives and friends wanted to bury them but could not identify the remains, Simonides was able to recall exactly where each person had been seated at the table. He realised the truth that 'the best aid to clearness of memory is orderly [visual] arrangement' (Cicero, *de Oratore* 2.86.351–54). In the circumstances a very composed thought!

In the late 18th century William Playfair, a Scottish engineer is credited with introducing the world to graphical representation of data in *The Commercial and Political Atlas* that contained line graphs and a bar chart. In 1801 he published *The Statistical Breviary* introducing the use of pie charts. It was a new way of encapsulating statistical data.





Later Florence Nightingale used pie charts to great effect to demonstrate the effectiveness of her methods to reduce mortality in the Crimean War.

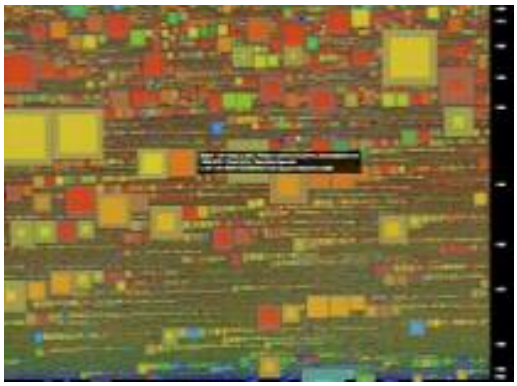
Manuel Lima states in his introduction to *Visual Complexity: Mapping patterns of information*, 'Information visualisation is widely used as a tool for understanding data—i.e., discovering patterns, connections and structure'.

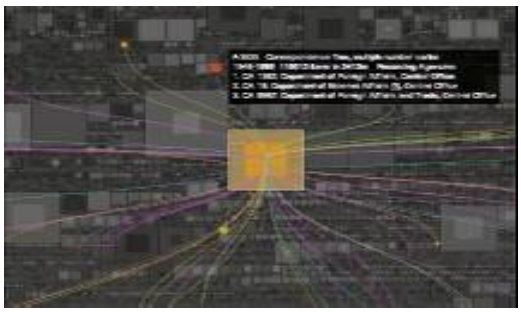
Eric Rodenbeck, founder of Stamen Design studio, speaks of data visualisation giving us new and unexpected ways to view data and draw unanticipated conclusions from it.

The National Archival and Records Administration in the United States has used the Texas Advanced Computing Center at the University of Texas to find solutions to the digital records challenge and concluded that:

> …to effectively respond to all of the requirements that are associated with very large digital record collections, innovative approaches and tools are needed…"
>
> Because large amounts of data cannot be comprehended at once, well-designed visualizations must provide a path that goes from an overview to a detailed perspective in order to facilitate a clear understanding of massive collections.

It was three years ago that I first became aware of the power of data visualisation. In 2009, Mitchell Whitelaw an associate professor in the Faculty of Art and Design at the University of Canberra undertook an Ian Maclean fellowship with the National Archives of Australia. He used visualisation techniques to explore the Archives' collections. In 2010 he demonstrated the results of his work to Archives' staff. This was an exciting revelatory moment for me, in which I saw the whole Archives' collection encapsulated in one space.





http://mtchl.net/visualising-archival-collections-the-visible-archive-project/

The image to the left is a representation of 65,000 records series held by the National Archives of Australia. It shows a timeline from 1900 at the top to the year 2000 at the bottom. Each square shows both a records series and the relative shelf space that it occupies.

In the image on the right you can see that by clicking on the descriptive metadata, relationships to other series are revealed. It is a visualisation that contains and shows everything, and has layers that can be mined to reveal more and more detail.

I found this ability to see the whole Archives' collection at once gave me a new appreciation of the extent, *shape* and relationships in the collection.

In his presentation to the ICA Conference in 2012 Mitchell Whitelaw argued a case for visualisation as a generous interface for collection investigation, rather than search which 'favours the expert user who understands a collection's contents and can query it effectively'. In spite of having vast numbers of digitised and digital records in collections, institutions still predominantly use search as the main tool to discover records. With ever increasing digital collections the main interaction with people is online. Indeed the National Archives' 2011–12 Annual Report that shows public reading room access and Commonwealth agency requests at 110,534. This figure is less than five per cent of the 2,303,992 online users of the Archives' collection.

He articulates five principles for the creation of generous interfaces:

- *Show first, don't ask*
  Search-based interfaces require a query to be entered, rather than presenting information about a collection.
- *Provide rich overviews*
  A visual interface should provide an interface that characterises the whole collection.
- *Provide samples*
  A visual interface should give cues that invite exploration.
- *Provide context*
  A visual interface should display the structure of a collection in a way that supports its interpretation
- *Share high-quality primary content*
  A visual interface needs to provide seamless access to high quality content.

The work described in Whitelaw's paper and his earlier work to visualise the archive show effective new ways to search and of seeing whole collections.

Lei Wang is a Senior Lecturer, at the School of Computer Science and Software Engineering, University of Wollongong. He has proposed content-based image retrieval for searching collections rather than the traditional word-based content retrieval. The approach that he uses is the 'Bag-of-Feature Model' whereby small patches of an image are strung together to form a visual word. Each image uses many of these visual words to create a visual dictionary. Each image can then use its own 'visual dictionary' to compare with other images and find correlations. This approach

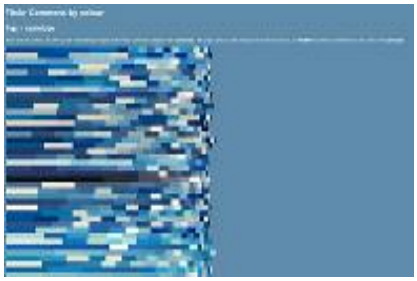
depends upon having a sample image to begin with. As a further step, Wang proposes a hybrid approach of using text-based search to initially identify the type of image required and then combine it with content-based image retrieval to produce a more effective image search.

From: Recent Developments of the Bag-of-Feature Model in Visual Recognition, 2012

Lei Wang's approach to visualising collections has great potential for accessing images and comparing their metadata to get a greater understanding of the image content and context. It is a tool for collection management and can confirm digitisation priorities. The Archives has continued to support the development of this image-based content retrieval system to enable the public to use it to search the Archives' photographic database.

Finally, Paul Hagon currently works at the National Library of Australia. In 2010 he undertook some interesting work where he used a tag to query images on Flickr Commons and display the representative colours of 50 images identified by that tag. By doing this he was seeing whether the colour gamut of particular processes could be found. He also generated a background that shows the average colour in the images.

This approach was a breakthrough for me – the concept of searching for images by colour gamut and the possibility of identifying processes.



Flickr Commons by colour tag: Cyanotype Flickr Commons by colour tag: 1970s

Paul Hagon's Flickr Commons project is an exciting step in using colour to identify palettes by process. As you can see by the 1970s images, there is a very familiar magenta cast to the images. To me this suggests a way of identifying groups of records that have deteriorated.

Additionally, he has developed a visual search tool that demonstrates the effectiveness of combining text and content-based searches. The example he gives uses an initial word search for Parkes. This can give at least three results – a person, a suburb of Canberra and a country town in NSW. The trick is that then by clicking on the word portrait to the left of the screen, a face recognition algorithm is initiated and the images of Parkes, the person are selected.

http://www.slideshare.net/paulhagon/everything-i-know-about-cataloguing-i-learned-from-watching-james-bond

Conclusion

This paper is a beginning. Further work developing a different approach in managing photographic collections will continue.

Physical surveys are traditionally used to examine collections and identify conservation issues and set priorities. They are usually limited by the availability of resources. Traditional database word searches dominate, but have the disadvantage that they give a blinkered view of collections and are limited by variability of the text descriptions.

Visualisation gives opportunities to harvest the vast amount of accumulated data and the vast numbers of digital images in collections and display them in new ways that allow unanticipated insights.

Mitchell Whitelaw's work demonstrates the ability to visualise large data sets at the one time. He also argues that visualisation, rather than traditional word-based

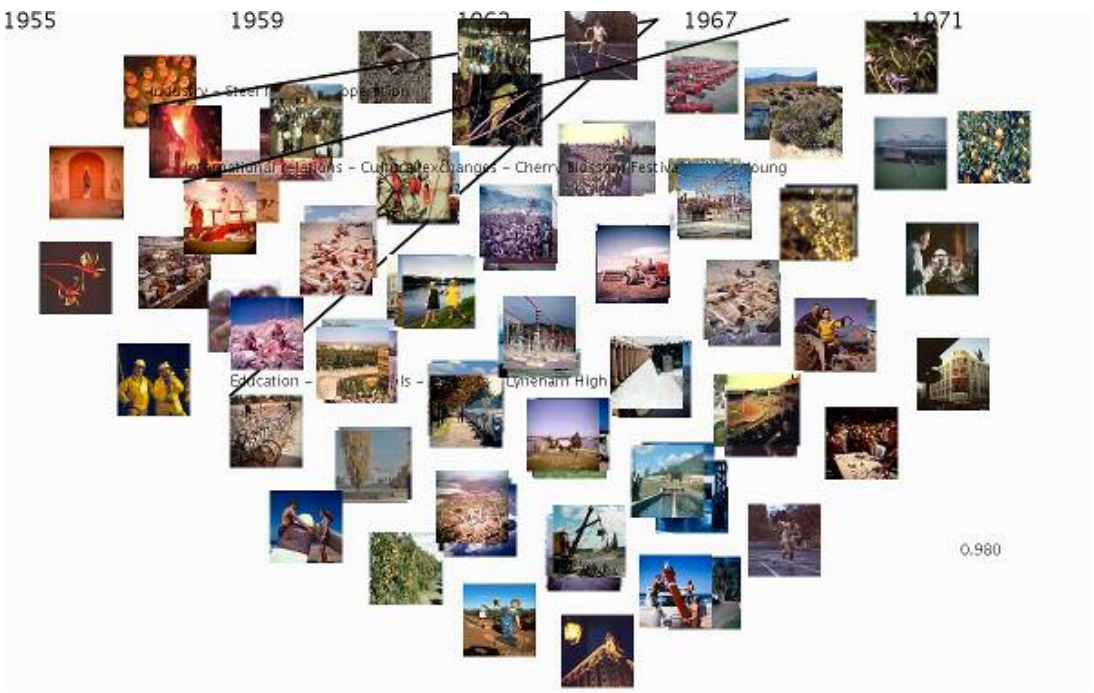
searches are more generous, because of its ability to give a sense of the whole collection.

Lei Wang's research shows new ways to explore collections and to link images and descriptive data in a way that can aid collection control. And potentially direct conservation effort. He also sees word-based search combined with content-based searches offering great potential.

Paul Hagon's work illustrates the potential of searches by colour and the power of hybrid word/image searching of databases.

I have made some explorations of data visualisation that are leading me to use colour to find deterioration patterns or identify particular photographic processes with known deterioration characteristics. My aim is to look at whole collections and use this approach to identify discrete parts that can be prioritised for preservation. The nature of digital design is that it is iterative and unexpected results or benefits can be achieved from experimentation.

I believe that we, as conservators, will be able to tap into this emerging approach to analysis of our large institutional collections and use it as an effective conservation tool.



This visualisation is a swarm sketch that has located the images according to their predominant colours. The lines from the images to the timeline indicate the date of each image's creation.

Author Biography

Peter Shaw joined the computer club in high school at a time when second hand computers were available for $25 a tonne. He now regrets ignoring computer programming completely in the intervening forty-five years. Peter came to conservation in the mid-1980s' graduating with specialisations in photographic and paper conservation. He has a keen interest in audio-visual and other electronic media and is the current coordinator of Electron, the AICCM's digital and audio-visual special interest group. In the spirit of personal growth coming from taking on difficult things, the paper he will deliver today has emerged from a growing fascination with the ubiquity of the digital world.

References

Cicero, MT, *de Oratore* 2.86.351–54, 55BC,
utexas.edu/research/memoria/Cicero.html

Hagon, P, *Everything I know about cataloguing I learned from reading James Bond*, 2010,
slideshare.net/paulhagon/everything-i-know-about-cataloguing-i-learned-from-watching-james-bond
paulhagon.com/commonscolour/
paulhagon.com/commonscolour/tags/cyanotype/
paulhagon.com/commonscolour/tags/1970s/

Lima, Manuel, *Visual Complexity: Mapping patterns of information,* Architectural Press, Princeton, 2011.

Playfair, William, *The Commercial and Political Atlas and Statistical Breviary,* London, 1786, Wainer, H and Spence, I (eds), Introduction, Cambridge University Press, Cambridge, 2005.

Roddenbeck, E, Stamen Design studio,
stamen.com/

Texas Advanced Computing Center, 2011,
tacc.utexas.edu/news/feature-stories/2011/a-window-on-the-archives-of-the-future/

Wang, L, *Recent Development of the Bag-of-Feature Model in Visual Recognition*, 2012,
http://www.naa.gov.au/Images/ICA2012PaperDownloadedVersion_tcm16-79232.pdf
see also, uow.edu.au/~leiw/share_files/Tutorial_SPS_School12.pdf

Whitelaw, M, *Visualising Archival Collections: The Visible Archive Project*, 2009,
mtchl.net/visualising-archival-collections-the-visible-archive-project/
see also, visiblearchive.blogspot.com, vimeo.com/6694353

Whitelaw, M, *Towards Generous Interfaces for Archival Collections*, 2012,
mtchl.net/towards-generous-interfaces-for-archival-collections/