

Digitise this: converting content

Cathie Jilovsky
Chief Information Officer
CAVAL Ltd
cathie.jilovsky@caval.edu.au

George Panagiotidis
Lead, Digital Services
CAVAL Ltd
george.panagiotidis@caval.edu.au

Janette Wright
CEO
CAVAL Ltd
janette.wright@caval.edu.au

Abstract:

This paper describes and illustrates the new processes that CAVAL is using to convert content into digital form. The research and development process began with the purchase of a Kirtas 2400 RA Book Digitising device in 2008. This Kirtas with its page-turning technology now underpins the provision of high quality digitisation services for books and bound volumes. Issues discussed include image enhancement, file format options, differing criteria for preservation and digitisation and the potential integration of complementary services such as metadata harvesting and copyright permissions management. More recently a facility to digitise large format newspapers and maps has been developed. Much of the digitised content created has now been made available via the web providing access to full-text searchable information that was previously almost impossible to find and access.

Introduction

CAVAL is a library consortium based in Melbourne, Australia, which has been providing shared services to member libraries and other customers in the region for over 30 years. The members include university libraries in Victoria, New South Wales and Tasmania, and customers are located throughout Australasia. As a not-for-profit organisation, CAVAL offers specialised services to support libraries and their constituent communities. This paper describes and illustrates some of the processes and techniques that CAVAL has developed to convert content into digital form.

A key service is the management of the CARM (CAVAL Archival and Research Materials) Centre, a purpose-built, high-density, environmentally-controlled storage facility designed for the long-term storage and preservation of print materials. The Centre has now been in operation for over 12 years and is nearing capacity. To meet future demand for the storage of low-use paper-based research materials, CAVAL has commenced building a second storage facility, known as CARM2.

Materials held in the CARM Centre shared collection are available through the Interlibrary Loan network. From its inception, the electronic copying of items, which may be rare and fragile, has been undertaken with due care. Initially, a Minolta face-up scanner was used to digitise pages, and each page was turned by hand. By the mid 2000s, the technology was outdated and this scanner was no longer supported by Minolta.

In 2007, CAVAL began the search for a replacement scanner, a key criterion being that it would be suitable for rare and fragile items. Following a market review and a careful investigation of the alternatives, automated page-turning technology patented by Kirtas was selected as the replacement digitising device.

The Kirtas 2400

The Kirtas 2400 RA Book Digitising device (Kirtas, 2009) was delivered in mid 2008. This device is similar to that used for the Google and Microsoft Book Digitising Projects and can digitise pages from bound volumes at up to 2400 pages per hour, capturing the contents of each page using two 21 Mega Pixel Canon Cameras. The bound books are held in place by a specially designed cradle, holding the book at an opening angle of 45 degrees. A series of lasers and processors control the robotic arm which turns the pages of the books. Air is used to gently lift and turn the page. Optical Character Recognition (OCR) technology is used to convert images into text.

The acquisition of the Kirtas 2400 RA not only provided CAVAL with the ability to continue to digitise CARM materials requested via Inter Library Loan, but also added the capacity to develop a digitisation service for members and the wider library community.

Following initial training provided by the supplier, the first task was to review the existing inter library loan workflows and to adapt them to the new technology. As part

of this process, several staff members gained practice with operating the Kirtas equipment. Staff were encouraged to experiment with the technical settings of the Kirtas as well as to become familiar with the associated software. The cut-over to live operations for interlibrary loans from the CARM Centre was achieved in July 2008.

Processes and workflows

The next task was to develop, test, document and implement procedures and workflows for the digitisation service.

The component tasks - document capture, automated batch processing, quality assurance, creation of output files in PDF format and the creation of searchable text using OCR – were firstly defined and documented. Training was the next step and this proved to be straightforward. Staff, many of whom are *Gen Ys*, were often able to learn the entire process within a day, a not surprising outcome given that they grew up with laptops, mobile phones, the internet and iPods. After three months, most of the software ‘quirks’ were found and rectified. Although each digitisation project is different, the general workflow structure is the same.

Unlike other technology implementations undertaken by CAVAL, we have found digitisation is not a ‘set and forget’ type of system. Improving our workflows and testing alternative software have become continuous tasks. Sometimes the catalyst is our client base, but often, it is our own *Gen Y* staff asking ‘but why can’t it be done better?’ Indeed, a healthy competitive culture in CAVAL’s digitisation department drives our Research and Development process. Research and Development is now a regular part of our digitising budget. An unplanned result of this is the maintenance of positive staff morale. Research and development tends to break the monotony of the job, and promotes staff ownership of the processes.

Early in the implementation, a process management system was overlooked. A series of complicated spreadsheets were used to keep track of each item, and how long each process was taking. At the conclusion of CAVAL’s first major digitisation project, (around 120 books), a proper process management system became a high priority. Building on this experience, we were then able to design, program and implement a dedicated database capable of tracking each item and project. This process management system is now integral to our operations, and allows the operational staff to plan, and distribute workloads with little (and often without any) management supervision.

The development and documentation of workflows and processes was an integral component of setting up a high quality digitisation service for books and bound volumes. More recently, we designed, developed and constructed a device that allows digitisation of large format newspapers and maps. Initially this device catered for items up to A0 size only, but a later modification removed this restriction.

Now, 12 months later, CAVAL is digitising a wide variety of material such as theses, university calendars, council minutes and high school chronicles. The common characteristic of these collections is that they must all remain in their original condition.

Digitisation technology

Prior to the announcement of the Google Books project in December 2005, efforts by individual libraries to digitise books were 'slow, expensive and underfunded' (Rouse, 2006). The Google project pioneered the development of new technologies, including the scanning of pages of brittle old books at high speed without damaging them, and managing huge amounts of data.

Karen Coyle (Coyle 2006) outlines three categories of digitisation projects -

(i) mass digitisation: the conversion of materials on an industrial scale. This has been made possible through the technology improvements: photographing books page-by-page and using OCR to produce searchable text.

(ii) non-mass digitisation: the careful and individual selection of materials to be digitised, along with richly marked-up text that can be used to provide a variety of services.

(iii) large-scale digitisation: more discriminating, these projects are concerned with the creation of collections and complete sets of documents e.g. JSTOR. (JSTOR, 2009)

Earlier scanning technology involved either flattening books onto a face-up scanner or removing the bindings so that individual pages could be fed through a document feeder. These 'read and destroy' digitisation projects have continued where a decision has been made that retaining the physical items is unnecessary, or where other copies of the books are available. Document feeders are relatively cheap to purchase and operate very quickly. For example, fifteen thousand images per hour can be achieved by a small digitisation operation.

Comparing this to the Kirtas page-turning technology of around 2,400 images per hour, it is clear that digitising using document feeders will inherently cost less 'per page' than page turners. However, document feeders are not suitable for rare and fragile items, as disbinding is often not an option. So, where items need to be preserved and digitised, the choices are automated page turners, manual face-up scanners, photocopiers and specialised photographic studios. Face-up scanners can digitise about one thousand images per hour, whereas photocopiers and photography studio outputs are slow and expensive.

Therefore 'read and destroy' projects will be most efficiently and cheaply undertaken using document scanners, and 'preservation' projects will be most cost-effectively handled by automated page turners.

In reality, page-turning technology is still in its infancy, and hence, the devices are still relatively expensive. Page-turning technology is currently about sixteen times slower than document feeder technology. In the near future, with post-capture software improvements, page-turning technology may be only eight times slower than document feeders. The release of a 3000 image per hour page-turning device by Kirtas in October 2009 further closes the gap between document feeders and page turners.

Physical limitations currently prevent automated page-turning technology operating as quickly as document feeders. Research into increasing the speed of page-turning devices is ongoing, such as that at the University of Tokyo Ishikawa Komuro Laboratory. For example, researchers there recently announced the development of a device that can scan a 1,000-page book in four minutes (Nakashima et al 2009).

Although scanning is clearly a key part of the digitisation process, it is only one part of the workflow. The other components, which will typically include the creation of metadata, image processing, quality control, OCR processes, creation of technical metadata, storage, data compression, indexing and making the outputs available through a user interface, will each take time and resources. It should be noted that the scanning is rarely the bottleneck, it is usually budget issues, rather than scanner speeds, that constrain digitisation programs.

Accuracy rates of 98 percent, or even greater, can now be achieved with OCR software. However, this is very dependent on the physical quality of the item being scanned. Note that 99.9 percent accuracy means that 1 character in 1,000 is wrong, averaging one error per modern book page (Coyle 2006).

Several recent articles in the library literature describe methodologies and experiences with digitisation. The San Francisco Public Library reported that 'mechanics and labor were more time consuming than anticipated' (Goldstein, 2009). This included quality control, metadata verification and network storage. Criteria for the selection of items to be digitised were based on usage, so that 'not the collection jewels, but the work horses' were prioritised. The University of Maryland focused on expanding existing workflows and expertise within the organisation as a way of incorporating digitisation into core library functionality. In particular, items were digitised in response to patron requests and added to a newly created digital repository, thus prioritising access to preservation (Gueguen and Hanlon, 2009).

The essay by OCLC Programs and Research, reporting on a 2007 Forum 'Digitisation Matters', explores strategies for scaling-up the digitisation of primary materials in light of the expanding book mass-digitisation programs (Erway and Schaffner 2007). Libraries are urged to build programs, rather than projects, through the integration of digitisation into workflows and user services, and to ensure that the content is exposed to search engines and aggregators.

Collaboration and Innovation

Page-turning digitisation devices such as the Kirtas 2400 are expensive to acquire and implement. Incorporating the additional but essential costs of staff, training and project management puts the total cost of owning this technology in-house out of the reach of most individual libraries.

CAVAL has a long history of providing innovative services for the Australian library sector and is uniquely placed to make investments in new or developing technologies specific to the library industry, the Higher Education sector and also of interest to other cultural institutions such as galleries and museums.

Unlike the 1970s and '80s, when VALA and CAVAL were established, we now operate in an environment in which a greater number of technological research and developments are destined for commercial exploitation. The Higher Education sector in Australia (as elsewhere) is more competitive and less inclined to invest in technology for the 'greater good' of the sector. Slaughter and Rhoades outline in their book 'Academic Capitalism and the New Economy' that the dominant model for the transmission of technological developments to the wider community is now the market model (Slaughter and Rhoades 2004).

As a non-profit consortium, CAVAL has the infrastructure to aggregate the demand for specialised services to achieve economies of scale. Examples include the introduction of the Turnitin (Turnitin, 2010) plagiarism-detection software; the development of the Virtual Document Exchange (VDX) ISO-compliant ILL (inter-library loan) software; specialist language cataloguing and large scale shared off-site storage facilities.

At the time when Kirtas brought its 2400 per hour page-turning scanner onto the market in North America, CAVAL was in a position to assess the likely demand in Australasia for digitising content from bound volumes (journals and monographs). Having done so, CAVAL was able to risk investment capital and staff capability to be the first organisation in Australasia to acquire the technology and to effect a 'technology transfer' to the region. As can be seen in our later discussion of transport risks, location is a key element of the success of a digitising service for preservation.

Collaboration by groups of libraries facilitated by organisations such as CAVAL, which can offset the costs with large volumes, provides a way to access this digitisation technology. In the future, as the technology improves, and equipment prices fall, then these devices may become standard library equipment.

Digitisation Services

A key feature of the CAVAL Digitisation Service is the Quality Assurance process including the cleaning and cropping of images.

Each individual page is checked 'on screen', by human eyes, for legibility, focus and cropping. This process is completed *before* the books are returned to the respective library, to ensure that, should quality issues be detected, individual pages, or, occasionally, entire items can be recaptured whilst on-site at CAVAL.

Cropping images accurately is inherently difficult for page-turning digitisation devices. As the pages of the book are turned, the left side of the book becomes thicker, and the right side becomes thinner. The Kirtas device does make provision for this correction, through the horizontal adjustment of the book cradle to keep the book centralised to the cameras; however, it is not perfect.

Further research has identified areas of the quality assurance process that can be assisted via the use of software, in particular automated and accurate software cropping. We are hopeful that significant reductions in the staff time required for quality assurance are achievable.

The cost of staff to supervise the capture and to manually quality check is currently the largest component of the total cost for any digitisation project. One possible option would be to move the manual quality assurance processes off-shore to locations with lower staff overheads. This would significantly reduce the staff-related costs, but add to the delays in ensuring the effective capture of the content before releasing the original materials from the CAVAL site. Several alternatives are under investigation.

Feedback from existing members and clients is that shifting the entire process offshore would be viewed as an additional risk and defeat the purpose of CAVAL offering this service in Australia. Given that the digitisation of rare and fragile materials is a risk management exercise, many libraries feel that transporting the physical materials offshore would be an unacceptable risk. However, CAVAL clients appear to be more comfortable with the option of capturing the digital asset on-site and outsourcing the manual Quality Assurance processes to an offshore organisation under CAVAL's supervision.

Risk management

Digitising rare and fragile books can be risky. Items may be damaged or lost during transit, or may be damaged during the digitisation process itself. By October 2009 CAVAL had captured over 220,000 pages and its Kirtas device, located in Melbourne, has not torn any pages, or damaged any books.

A variety of transportation methods have been used to send items to CAVAL for digitising. We have seen the usual standard Australia Post, Express Post and a great variety of couriers. We have also had items arrive with the client 'in-person' via hand luggage on a domestic flight, and via TNT's failsafe service. The latter is a very expensive service that ensures the box containing the rare and fragile books is sent to the destination as quickly as possible, and guarantees that the box is treated with care.

Some organisations have identified an important collection which would benefit from digitisation, but have not come to terms with transporting the items for digitisation. For valuable collections, some libraries deem that the risks involved with transportation outweigh the digitisation benefits. The use of specialist courier services that can mitigate the transportation risk, or, for a low number of volumes, personal delivery by a staff member may be acceptable options.

Another possibility we have considered is for CAVAL to arrange a Kirtas Roadshow, potentially visiting each Australian state with the Kirtas equipment and trained staff. The equipment could then be set up in a suitable host library and the rare and fragile material from surrounding institutions could then be digitised. This would reduce the transportation risk from interstate distances to local distances.

Some libraries assume that the Kirtas page-turning technology is a hazard for rare and fragile books. Generally, a live demonstration of the equipment is convincing, often using the prospective client's sample materials. The Kirtas uses air to gently

separate the pages of the book, and an automated arm uses a gentle vacuum to lift and turn the page. Several demonstration videos are available on 'YouTube'.

Output formats

Given the stringent digitisation risk assessments for rare and fragile items, it is especially important that the digitised item not be subject to the same risks ever again. In order to future-proof the digitisation of rare and fragile materials the system should be organised so that items are captured only once. The future-proofing methodology we have developed is fairly simple: regardless of the output required, full colour, greyscale or black and white, everything is captured in full resolution (using 21 mega-pixel cameras) and full colour (24 bit). We then down-sample, quality assure and convert to the required format (usually PDF).

As well as providing the final format, we also provide the original full resolution and full colour images. The original image is a by-product of the page-turning technology and there are no additional costs involved with the provision of the images. If at any time in the future, the PDF (or other chosen format) is no longer suitable, the original images can be reprocessed, and re-formatted without a loss of quality. Alternatively, the original images may be reprocessed into different formats, such as XML. It is not good practice to reformat down-sampled images, as each re-format reduces the resolution. And most importantly, the future-proofed images can be reprocessed, cheaply, without having to re-digitise the physical item and subject it to additional risk. This is dependent on appropriate protection and management of the digital files.

Digitisation projects fall into two categories, preservation and access. Access type projects require the file size to be as small as possible. Preservation projects are not governed by file sizes, and often the colour of the page is deemed almost as important as the content itself. The 'look and feel' projects are popular, accounting for 40% of all CAVAL's projects.

Most often, the output files are made available to the general public via the internet, for which the right file size is a critical factor. Usually, a smaller compressed file, heavily down-sampled, will be the most suitable for on-screen view. When printed, these files are legible, but high quality printing is not possible. Down-sampling and compressing images for internet transfer does have a negative impact on image quality, however, the small overhead associated with storing the full resolution, full colour future-proofed images will negate the need for expensive re-digitising.

We have had countless discussions with our clients about PDF file sizes. Black and white (bi-tonal) images produce the smallest file sizes, but the process discards colour and all greys. A handy by-product of bi-tonal images is that the background colour of the page is converted to white. This makes the text easy to read; however, it makes all colour and grey pictures illegible.

The software provided by Kirtas contains a useful feature known as 'segmentation', which allows images to be converted into bi-tonal format, but retains the colour and grey in the pictures, diagrams and graphs. Essentially, this gives the best of both worlds, enhancing text areas with white backgrounds, compressing file size, and

maintaining picture quality. The output provided by the segmentation function is ideal for items that will ultimately require re-printing. Ideally, items digitised for re-printing should contain a white background, instead of the dull colour of the original page. Segmentation not only provides the white background, but also maintains picture, diagram and graph quality required for printing.

XML (eXtensible Markup Language) is another format emerging as a requirement for CAVAL clients. Unlike the previously mentioned formats, XML is not a straightforward or automated process. The extraction of the text from the images via OCR is reasonably effective, with a 96% accuracy rate on average. The next step is then to 'arrange' the text into a predefined XML format. Once formatted, the text is then tagged, for example, the beginning of chapters, placement of pictures/diagrams, identifying page numbers, etc.

Manually tagging the XML file is very time consuming and hence very expensive. CAVAL is currently testing another Kirtas software application, which enables automated XML conversion. There are scores of XML schemas to choose from, and our observation is that most libraries are not entirely aware of the full extent of XML requirements. There is a trend to 'digitise now, XML later'. Provided the full future-proofed images are kept, XML conversion can be easily undertaken at a future time.

Copyright

The last of the digitisation caveats is of course, copyright. Interestingly, we have not yet encountered many materials that are protected by copyright. We suspect generally that digitisation projects requiring page-turning technologies will not encounter many copyright difficulties, because this material is likely to be rare and fragile. Almost by definition, rare and fragile materials are usually old and out of copyright.

However, the introduction of the Digitisation Service did require CAVAL to consider this element. With some professional advice from the RMIT University Copyright Management Service, the CAVAL Digitising team was able to document a checklist of conditions which would require the client to seek Copyright Permissions where appropriate, and to indemnify CAVAL where the content was not in copyright. This led to a more formal process for quoting and accepting prices and service conditions for the CAVAL Digitisation service.

Complementary Services

As a result of being the first organisation to implement this data capture and digitising technology within Australia, CAVAL has begun to consider a number of potential complementary services including a commercial Copyright and Permissions Service and the delivery of digitised content on a desk-top e-book platform. Some of these initiatives are under discussion with potential partners and the CAVAL Copyright and Permissions Service will be launched in 2010.

Conclusion

Our research and subsequent production experience has confirmed that page-turning digitisation technologies are reliable and the most efficient at digitising rare and fragile materials. Some improvements in post-processing methodologies will reduce the costs associated with the quality assurance process, but the physical limitations of page-turning technologies are not expected to allow for speeds comparable to that of document feeders.

Librarians and archivists are becoming more comfortable with digitising rare and fragile items using page-turning technology, and are mitigating transport risks with alternate and safer transport options. Libraries are digitising 'smart not hard' and ensuring expensive digitisation projects are future proofed. Digitisation and preservation need not be mutually exclusive, as by using page-turning technology it is no longer necessary to disbind bound items for digitisation.

Since the implementation of the digitisation devices and subsequent research and development, several hundred thousand digital images have been produced by CAVAL. Most of these processed images have been made available via the web providing access to full-text searchable information that was previously almost impossible to find and access.

The acquisition of sophisticated digitising equipment and software, and the development of associated techniques, workflows and processes has been both exciting and challenging for CAVAL staff. It is an example of the way in which CAVAL, representing its members in the Higher Education sector and the wider library and archive community, and with sufficient infrastructure, can invest in and develop technologies and services for the local industry. .

References

Coyle, K 2006, 'Mass Digitisation of Books', *The Journal of Academic Librarianship*, vol 32, no 6, pp 641-646.

Erway, R & Schaffner, S 2007, 'Shifting Gears: Gearing Up to Get Into the Flow', Report produced by *OCLC Programs and Research*, viewed 26 October 2009, www.oclc.org/programs/publications/reports/2007-02.pdf.

Goldstein, S 2009, 'The Good, The Bad, & the Ugly: Digitizing on a Shoestring', *Califa 2009 Digital Symposium*, viewed 26 October 2009, http://www.califa.org/cdm_slides.php.

Google Books Project, viewed 26 October 2009, <http://books.google.com/googlebooks/library.html>.

Geugen, G & Hanlon, AM 2009, 'A Collaborative Workflow for the Digitization of Unique Materials', *The Journal of Academic Librarianship* vol 35, no 5, pp 468-474.

Kirtas, viewed 26 October 2009, <http://www.kirtas.com>.

Nakashima T, Watanabe Y, Komuro T & Ishikawa M 2009, 'Book Flipping Scanning', viewed 28 October 2009, http://www.k2.t.u-tokyo.ac.jp/members/watanabe/nakashima_uist09.pdf.

JSTOR, viewed 26 October 2009, <http://www.jstor.org>.

Rouse, W 2006, 'The Infinite Library', *Technology Review*, May 2006, pp 54:59.

Slaughter, S & Rhoades G 2004. *Academic Capitalism and the New Economy*. John Hopkins University Press, Baltimore.

Turtinin, viewed 19 January 2010, <http://turnitin.com>.