Ramping it up: 10 Lessons learnt in mass digitisation

Rose Holley*

In 2007 the National Library of Australia (NLA) began a large-scale newspaper digitisation program that aimed to digitise one million pages (10 million articles) per year, with a view to increasing the volume over time and ramping up digitisation to include books and journals as well as newspapers. By the end of 2009 the NLA had learnt 10 key lessons about ramping up its digitisation activities into a mass-scale operation.

BACKGROUND

The National Library of Australia (NLA) took an early lead in digitisation activities in the 1990s and by the year 2000 had established a solid digitisation policy that was publicly available. Digitisation moved rapidly from project based to normal business, with a methodical approach to digitising the NLA's collections in-house in the state of the art scanning studio. By 2008, over a 10-year period and with the assistance of 20 full-time staff, the NLA had carefully digitised to a high standard 100,000 items from its collections.

In 2007 the NLA embarked on its first large-scale digitisation program. This was the national collaborative Australian Newspapers Digitisation Program (ANDP), which aimed to digitise one million pages (10 million articles) per year.[1] This is not necessarily mass digitisation, but rather large-scale digitisation. However, it was the clear intent of the NLA to use this experience to ramp up digitisation activities in the coming years; both by increasing the volumes and the types of items to be digitised. It was anticipated that by 2010 the NLA would have the capability and knowledge to be undertaking mass digitisation of books, journals and newspapers. If the capability was there, the NLA could just focus on getting the funding.

It was decided to outsource large-scale digitisation work, as no attempt would be made to set up an in-house mass digitisation operation. The existing in-house, high-quality digitisation operation continued with a staff of 20. The new mass digitisation team consisted of five full-time staff. Of these, two were managers (managing processes, workflows, contracts, software), the third was in charge of logistics (movements of materials) and quality assurance, and the remainder two positions were comprised of 10 part-time staff undertaking quality assurance work.

Much has been written about large-scale and mass digitisation issues generally, so this will not be covered here. Instead, the focus is purely on what the NLA has learnt so far about ramping up digitisation into a large-scale/mass digitisation operation. The NLA reviewed what had been learnt after three million newspaper pages had been digitised over three years. Although the lessons apply to newspaper digitisation they would equally apply to journal or book digitisation. The NLA wants to share the lessons learnt more widely so that State and Territory libraries in Australia and other national libraries can confidently move forward into mass digitisation with as much knowledge as possible.

MASS DIGITISATION ISSUES ADDRESSED

The NLA had a good understanding of current mass digitisation issues before the "ramping up" began, as these were taken into consideration and addressed before the ANDP commenced in 2007. Some of the current issues at that time and their solutions are outlined in an essay by Ricky Erway and Jennifer Schaffner entitled *Shifting Gears: Gearing Up to Get Into the Flow*.[2] The essay was inspired

[*] Manager, Australian Newspapers Digitisation Program, National Library of Australia.

[1] Australian Newspapers Digitisation Program, http://www.nla.gov.au/ndp viewed 8 January 2010.

[2] Erway R and Schaffner J, *Shifting Gears: Gearing Up to Get Into the flow* (OCLC Programs and Research, 2007), http://www.oclc.org/programs/publications/reports/2007-02.pdf viewed 8 January 2010.

by the *Digitization Matters* Forum, where 200 organisations were invited to focus on ideas for significantly increasing the scale of digitisation activities in libraries. The eight key ideas discussed in the essay and the actions the NLA decided to take around these issues/ideas are described below.

Access versus preservation

Before the digitisation began it was agreed this was an access not a preservation objective.

Selection has already been done by users

The process for selection was made clear to stakeholders and the public, and the emphasis was on content to which the public wanted access (based on newspaper microfilm usage, sales figures and suggestions).

Do it once (then iterate)

Because of the uncertainty of which image file (greyscale or bi-tonal tiff) would be of the most use in the future for OCR (optical character recognition) and also public delivery, a matching pair of files was output for every page so that the NLA would not have to go back and digitise again. Either file could later be manipulated, enhanced or derivatives created. OCR contractors could not easily or effectively process jpg files so tiffs were created, despite the fact this was an access not preservation program.

Programs not projects

The ANDP was called a "program" rather than a project right from the start. This helped to convey both the large-scale and long-term digitisation intentions of the NLA.

Description: You do not need a catalogue record or detailed description first

Many of the newspapers did not have a catalogue record. This would not be a barrier to digitisation and would/could be done later. It was acknowledged that the public could help in description of content, eg identifying incorrect page numbers or adding keyword tags to articles.

Quality versus quantity

It was quite clear that the public would prefer to have as many pages of a newspaper as existed rather than a few high-quality pages. It was decided to deliver all pages from the microfilms no matter what the image looked like (even page fragments), and to undertake basic image quality spot checks (eg "Is the page the right way up?") rather than detailed image quality checks and manipulations (as existed for in-house digitisation).

Discovery happens elsewhere

When the search and delivery system was built it would be indexed by Google, have links into Wikipedia and harvesting sites would be set up for any organisation that wanted them. This was of high importance since it was decided the data should be made as accessible as possible. It was assumed from the start that the discovery path for most users would be via Google or Wikipedia, not the NLA website.

Brother can you spare a dime?

Although there was a $10 million budget, this would only cover a fraction of the content scope. Organisations and individuals who wanted content included would be offered a model of sponsorship or contribution, so that the cost could be shared. Access would always be free to all via the internet.

Thinking upfront, making these decisions and then sticking to them helped to eliminate many of the issues that other organisations have faced in attempting to scale up their digitisation activities. The decision to outsource digitisation (at least for the first four years) was largely based on the fact that there was not adequate or suitable space within the NLA (a historic building) to accommodate a large-scale digitisation operation.

Other obstacles to mass digitisation, as outlined in a paper by Astrid Verheusen of the Dutch National Library – Koninklijke Bibliotheek (KB),[3] that were addressed are:

Technical infrastructure

The existing infrastructure would not be able to cope with mass digitisation. A new technical infrastructure would therefore be created. This would initially only be used to test the ANDP. It would include backend functions and digitisation workflows, and also a public search and delivery service. This infrastructure would have scalability, and long-term would replace all the existing multiple NLA infrastructures. Mass digitisation of any type of materials would be able to use the new infrastructure; and delivery of all new services would use the new delivery architecture. This model was referred to by the NLA as the "Single Business" project and was expected to take several years to fully implement.

Storage

A massive amount of storage (both online and offline) would be needed. Therefore a healthy budget would be allocated for storage and additional staff resource would be allocated to implement and manage the expansion of the storage system. As far as possible, movement of files would be automated.

Quality assurance

Quality assurance would be undertaken but as far as possible this would also be automated, eg on data ingest check directory structures and file formats. Manual quality assurance would be based on ISO sampling techniques which would involve checking about 3% of images. The checking would be for basic things such as image orientation rather than image quality.

Project management versus normal business

Since this was the first large-scale digitisation the NLA had undertaken, there would be a "project phase" which would involve setting up and implementing the new infrastructure and storage, software development of new systems (search and delivery and content management), and selection and setup with digitisation contractors. This was estimated to take around two years and would be managed by an experienced project manager using PRINCE 2. At the end of two years, large-scale digitisation would become normal operational business for the Digitisation and Photography Branch at the NLA. It would not be treated as "project work" or managed as a "project", but rather as day-to-day, normal business utilising regular digitisation staff.

The 10 Lessons Learnt

Despite coming to mass digitisation with a solid body of knowledge, firm plans, and risk management strategies in place, the NLA learnt lessons in mass digitisation (both good and bad) that will help in the current endeavours to "ramp it up" even further. The NLA reviewed what had been learnt after three million newspaper pages had been digitised over three years and these lessons learnt were tabled at the project closure meeting in 2009, where the Director-General of the NLA decided they should be shared more widely within the library community.

Storage

The amount of online storage needed was initially underestimated, and online storage could not be purchased or implemented easily within the timeframe required. It took several months (due to Request for Tender (RFT) procedures, training requirements, and waiting for hardware deliveries from overseas), which resulted in temporary work around solutions being implemented, none of which was very effective. Storage issues caused the workflow to frequently slow down, stop or be significantly changed, which had a major impact on the program.

[3] Verheusen A, "Mass Digitisation by Libraries: Issues Concerning Organisation, Quality and Efficiency" (2008) 18(1) *Liber Quarterly* 28, http://www.liber.library.uu.nl/publish/articles/000225/article.pdf viewed 8 January 2010.

The amount of online storage required could not easily be estimated because it was not known how closely the contractors would work to the timelines and workflows given. The termination of a contract with an OCR contractor caused a two million page backlog and resulting storage issues. It had been anticipated that the workflow would involve a continuous process of scanning then OCR. However, in actuality, the workflow became divided into two separate parts, with months and sometimes years between each process. This resulted in millions of files needing to be moved offline and then later recalled online, which was a major change to the planned workflow and storage. It also necessitated changes to the file naming and renaming process. The moving and copying of files took excessively long (days or weeks). This kind of time to move files had not been accounted for in the project plan. In addition, for the majority of the project phase the file moving and copying was largely a manual process, using vast amounts of a programmer's time and requiring regular manual intervention when error messages occurred. Two months before project end most of the file moving was finally automated, and could be instigated by digitisation rather than by IT staff, but it had taken much longer than expected.

Although there was plenty of money for storage, the IT department was reluctant to spend the money until more storage was absolutely necessary, because storage prices were dropping all the time. Rather than buy the storage upfront as the digitisation team had expected, it was purchased in small lots several times, which delayed processes, full automation and final implementation. Communication between the IT purchasing staff and the digitisation staff was unclear and the interpretation of "coming soon" was different to each group.

Lesson learnt 1: Storage

Storage should either be purchased upfront when the business area first requests it so that workflow and project timelines are not impacted, or there should be a clear understanding between the IT department and the business area of what the order to implementation finish time actually is, so that dropping storage prices can be taken advantage of without project timelines being affected. In a large-scale digitisation program it is better to assume that digitisation contractors will not work to the timelines given and plan contingency and alternative storage arrangements upfront. Automating storage workflows is challenging and time consuming.

Better systems for the business area to view, manage and monitor storage were required. Not having access to storage usage information and only being alerted when no online space was left made managing the workflow difficult.

Quality assurance

Most of the key quality assurance (QA) processes were automated. However, manual QA was undertaken on both the scanning and OCR work (based on ISO sampling). This quality assurance work was quite basic and not to do with image quality but rather checking that files met digitisation specifications. A few seconds per image were allocated and QA staff could work through thousands of images quite quickly.

Although the new mass digitisation team were comfortable with this, this was not fully embraced by the in-house digitisation team, who were more accustomed to performing true image quality checking and enhancement on individual images with no time limit (often taking several hours per image).

Lesson learnt 2: Quality assurance

Some parts of the mass digitisation workflow, especially QA, are considerably different to those of high-quality individual digitisation and this needs to be both acknowledged and embraced by managers and operational staff. Digitisation staff may find it difficult to transition between the two types of QA effectively and therefore two QA teams may be required (one for mass digitisation and one for high-quality individual digitisation).

Quality assurance staffing

Due to the repetitive nature of the quality assurance work, 10 staff worked in single three-hour shifts. It was not feasible for staff to work full-time or full days on quality assurance work. The amount of incoming QA work was variable and sometimes unpredictable and was therefore hard to plan in advance. QA work had to be done within a certain timeframe as specified in the digitisation contracts, so when work came in it had to be checked as soon as possible. Taking all these factors into account and the high level of flexibility required, utilising existing in-house digitisation staff was impractical.

An alternative scenario was implemented and was highly successful. This was the appointment of casual student workers who gave the mass digitisation team the flexibility it required for staffing, and high quality, keen and enthusiastic staff members. However, the appointment process was costly and time consuming (being the same as that for ongoing public servants in government institutions) and did not give the ANDP team the option to quickly appoint staff as was required. After a year it was agreed that the NLA would implement a more cost-efficient and speedy method of recruitment for casual ongoing staff, on a new form of contract better suited to mass digitisation activities.

Lesson learnt 3: Quality assurance staff
Mass digitisation requires a high level of flexibility in QA staffing, due to the variance of workloads and reliance on incoming work from external contractors. It also requires speedy and cost-effective recruitment processes so that timelines and budget are not negatively impacted. It was beneficial that the HR manager listened to the needs of the business area and was able to adapt recruitment methods and contracts for casual ongoing staff to meet the needs of mass digitisation work. The NLA must remain willing to review particular staffing needs for mass digitisation in the future.

Digitisation contractors

It was initially thought that one contractor could do the entire digitisation process, but it quickly emerged that capabilities required for scanning and OCR were quite different and that contractors preferred to do one or the other, but rarely both. Therefore one scanning contractor and one OCR contractor were appointed. However, this resulted in high levels of risk when one of the contractors failed to deliver within agreed timeframes or to agreed specifications. The contractors had dependencies on each other and the NLA had dependencies on both. Due to a termination of one of the contracts, the government RFT process had to begin again. This time, to reduce risk, it was decided to appoint a panel of contractors. However, the time lag in the government RFT process and the follow-on contract negotiation process meant that it was in the first case two to three years later when the selected contractor was appointed, and in the second case one and a half years later when the contractors were appointed. The contractors were signed up for two to five years. The IT and digitisation contractor marketplace is quickly changing and evolving and time lags of this duration (because of following due government process) make working with the most ideal contractor at the right time unlikely.

Lesson learnt 4: Digitisation contractors
Appointing multiple digitisation contractors to a panel reduces the risk for mass digitisation if individual contractors cannot meet agreed specifications or requirements, or if they drop out of the marketplace. However, managing several contractors vastly increases the workload of the operations manager and the extra time must be accounted for in the project plan. The government RFT process is lengthy and there is a significant time delay between requests for tender, contract signing, and allocation of work. Adequate time, budget and staff resources must therefore be allocated to do this. In addition, the best contractor in the current marketplace may not be the one finally on the books because of time delays. Ideally, going out to market should happen every two years to make sure that the most cutting edge contractors in the marketplace are being used, although this is a significant drain on staff resources.

Digitisation contractors: Volumes

It was initially assumed that all the contractors would be able to meet the volumes, timelines and specifications as outlined in the RFT (to which they had responded to and agreed), and that they were aware that this was a mass digitisation project. The volumes set were at a level considered by the NLA to be the lowest for mass digitisation (one million pages per year), with room to "ramp up" work with newspapers and later to include journals and books as well.

Without fail, all five contractors that the NLA worked with in the period struggled to meet or could not meet the volumes within the required timeframes. This was a surprise to the ANDP team since most of the contractors had initially indicated that they wanted much higher volumes than they were given and could easily achieve higher volumes. Fortunately, the project changed from sole contractors to multiple contractors halfway through, which meant that one million pages per year could just about be achieved.

Lesson learnt 5: Digitisation contractor's volume

Although all digitisation contractors indicated they were equipped and ready to undertake "mass digitisation", in reality most were not. Spreading the work between several contractors vastly increased the workload of the project manager, but did achieve the minimum page target set per year (one million pages). It may still be some time before contractors can really achieve the mass digitisation volumes that libraries want. Contractors need to learn how to "ramp it up" as well as libraries.

OCR contractors: Setup and specifications

It was anticipated that project setup with OCR contractors would take around eight weeks. This was agreed to by contractors and written into the contract schedule. Contractors also agreed they were familiar with the ALTO and METS specification and were highly knowledgeable about zoning, categorisation and OCR. When the OCR contractor had not successfully completed a 50,000 page pilot to the agreed specification after 12 months and three attempts, it was thought that this was highly unusual. However, at the next IFLA Newspaper gathering, informal discussion amongst other national librarians undertaking similar projects indicated that this may be the norm rather than the exception. At least three other national libraries had experienced setups of similar periods with equally unsatisfactory results. Initial indications are that setup with any OCR contractor may take several months rather than weeks. It was surprising that OCR contractors seemed to have less, not more, knowledge than the NLA of zoning, categorisation, OCR, dictionaries, accuracy levels, relevant software and ALTO and METS specifications, despite this being their core business. The NLA found itself having to write several specifications and numerous iterations of processing and QA procedure documents for internal use by contractors in order to do the job. Several contractors requested to use the NLA's internal QA system since they did not have an adequate one of their own. The NLA had expected that contractors would be knowledgeable and helpful about newspaper and mass digitisation and that they would lead and advise rather than vice versa.

Lesson learnt 6: OCR contractor setup

The NLA has greater knowledge of all newspaper digitisation processes and software than any of the contractors on the existing panel, especially around OCR, zoning, categorisation and ALTO and METS files. It is likely that in the future as now the NLA will need to advise and guide contractors about issues, rather than vice versa. Setting up and undertaking a small OCR pilot is likely to take months rather than weeks. It is still unclear to the NLA why this is when all existing contractors already have considerable experience in this area.

Managing digitisation contracts

It was initially assumed that there would be good working relationships with contractors and that they could all meet the volumes, timelines and specifications as outlined in the RFT (to which they had responded and agreed). However, there were performance and management issues with all five contractors. Initially it was hoped to be able to amicably resolve these issues informally. While some contractors were customer orientated and were willing to listen and discuss the workflows and issues, others were not. A large amount of money was spent on legal advice and a large amount of managerial time was spent on issue resolution, ultimately resulting in a contract termination. Following this, the ANDP team were advised to take a "hard line" approach to contract management, rather than a "softly, softly approach", especially in regards to meeting timelines, specifications and volumes. Penalties were built into the contracts.

Lesson learnt 7: Managing digitisation contracts

Some digitisation contractors are not good at listening to or meeting customer needs and instead think the client should fit into their digitisation workflows and methods, which are set in stone. It is important to establish early on which type of contractor you are dealing with. It became apparent that the less flexible contractors were those who had less technology and skill. Therefore it is important to be working with a contractor who has an active R&D department, stays abreast of current software developments, and has a strong IT team. These are questions that can be asked in the RFT process.

There is a lot at stake in mass digitisation projects if contractors do not meet their volumes, timeframes or specifications. It can cause major impact on budget, timeframes, workflow and staffing, resulting in the entire mass digitisation program grinding to a halt and the public service being affected. Be prepared to implement contract penalties and to have a legal adviser and budget for the worst-case scenario. A business-like approach needs to be taken to managing contractors rather than a traditional ad hoc informal approach often taken by libraries towards their suppliers. Mass digitisation is a serious business.

Mass digitisation workflows

Establishing the workflow process and the Newspaper Content Management System (NCM) to support it was an ongoing and challenging process. The workflow process changed many times and currently three different workflows are going on consecutively in the NCM. The workflow process was initially evolving as the NLA considered and trialled various options. This was the first time the NLA had worked to the volume and scale of mass digitisation, and also the first time the ALTO specification had been used. The amount of time required to develop the NCM to meet the needs of newspapers, books and journals was underestimated, as was the scope of the entire system. There were several areas of workflow that had been discussed but not implemented in the NCM. This was because the NLA was leading the field with mass digitisation of newspapers and some ideas, eg post-OCR automated correction, had not yet been developed by contractors. In order to successfully ramp up mass digitisation and deliver innovative digital services to users it is essential to continually review and modify workflows and thus continue to develop the NCM. It is unlikely the system will come to a point of being "finished" anytime soon.

Lesson learnt 8: Mass digitisation workflows

The time to design a newspaper content management system was underestimated because it was not clearly understood at the start how complex the workflow process would be, that multiple contractors would be utilised, that different workflows would be undertaken in the system simultaneously, and that the system would later be required for all mass digitisation projects. It was not clearly understood by the IT team that the business area needed to constantly evolve a system to meet changing requirements in digital technologies and improved and new workflows. These changes often required (and still do) significant development work rather than just individual "enhancement requests". The NLA has learnt that mass digitisation workflows are highly complex and still evolving. The resource required to design a mass digitisation workflow and content management system is significant and ongoing. The workflows designed for newspapers are largely relevant and usable for books and journals with some minor changes required. However, mass digitisation workflow per se is still evolving and developing and will continue to do so into the future.

Transparent processes and progress

It is unusual for internal project management processes and project progress to be made publicly available by the NLA, but in this case the mass digitisation program was so high profile that, rather than a restricted internal wiki, it was decided to make most of the project documentation and progress reports available on a public website for both the public and the stakeholders. This was set up within a month of starting the program and was a wise move. Although the NLA knew the program was high profile it had not anticipated quite the volume of public interest that would be generated. The public website not only kept the public informed and stemmed some of the flow of enquiry emails about titles being digitised, it helped other national libraries that were working on similar projects, and assisted the stakeholders and also the contractors in understanding processes.

Lesson learnt 9: Transparent processes and progress

On a high profile mass digitisation program it is important to keep as many people informed with as much information as possible in a transparent way. It was a worthwhile idea to set up the ANDP public website for this purpose. It kept the public informed, helped other national libraries that were working on similar projects, and assisted the stakeholders and also the contractors in understanding processes. In addition, should any key staff have left, it served as a means of recording essential information in a single place. It also saved time long-term since much of the information on the site was requested multiple times for different purposes by different people; hence it made sense to summarise anything important at the time and write it up in a document for the website.

Level of public involvement

This was the first project the NLA had undertaken that involved mass public participation and interaction at several levels. This was enabled in several ways. Initially, the public were actively encouraged to sign up to become testers of the beta search and discovery system. It was not anticipated many would do this. However, within weeks thousands of users were "testing" the system so sign-up stopped and anyone using beta could be considered a "tester" – this finally amounted to half a million people. Once the public were testing/using the system, they were encouraged to give feedback and suggestions. This, again, they did at levels not anticipated. This amounted to thousands of requests by email. Subsequent to this, feedback was structured within surveys or restricted to small groups of users. Then public OCR text correction, tagging and commenting was implemented in the search service.[4] Again thousands of users began to interact with the service and create huge volumes of content. By November 2009, eight million lines of text had been corrected and 250,000 tags added by the public. This proved that users really did want to help the NLA and interact with data in new ways. At project close no marketing or promotion had been done for the service and no formal appeals for text correction made, yet despite this there was a significant and growing body of users.

After two years the public began asking how to get a title in the service. Many had asked if they could specifically donate money for digitisation of newspapers – either whole titles or page by page. Several sponsorships for titles had been received from organisations (without prompting), the most notable being a $1 million donation from the Vincent Fairfax Family Foundation to digitise the *Sydney Morning Herald*. It was the opinion of the project manager that if a "donate to digitise a newspaper page" was set up on the front of the service ($2 per page), a large amount of money could easily be collected to further digitisation of regional titles. A similar and very successful donation system had operated for Wikipedia, who received $6 million in donations (most under $10) from public users over a 12-week period. The public also suggested that they could make videos and put them on YouTube to promote the service if the NLA ran a competition for this.

Lesson learnt 10: Public involvement

The public responded in an unprecedented way to assist the NLA with its access and mass digitisation goals when given the chance. All sorts of opportunities arose to work with the public to achieve the goals (many of which have not yet been fully actioned). The assistance of the public should not be underestimated and is a powerful resource. What was surprising to many NLA staff was that the public should want to do this. It was not fully appreciated how keen the public are to help an institution so highly regarded as the National Library of Australia, what they can give us, and how desperately they want to interact and explore resources in new ways. Part of the public motivation to help lies in the fact that libraries are non-profit organisations and will not commercially exploit the volunteers' work, and it is obvious that libraries do not have the resource internally to achieve many of their goals at all or within a reasonable timeframe.

Ways in which the public can help include: correcting text to make the resource more accurate; adding comments to data to increase its value; tagging data to make it discoverable in new ways; donating money to assist in newspaper digitisation; creation of videos to market and explain the service; and helping to shape development of new services by giving feedback, especially those with Web 2.0 interactions. This is one of the most important lessons learnt and one which all other libraries should note. The public sincerely want to help and be involved as much as possible. We should take this opportunity and set an example that other libraries can follow.

CONCLUSION

The NLA has learnt useful lessons from the large-scale newspaper digitisation work that has occurred over the last three years. These involve storage, quality assurance, digitisation contractors and public involvement. The lessons learnt after digitising three million newspaper pages are equally applicable to book and journal digitisation. The NLA will be applying what has been learnt to existing large-scale digitisation operations and future mass digitisation operations. The NLA wants to share the lessons learnt more widely so that State and Territory libraries in Australia and other national libraries can confidently move forward into mass digitisation with as much knowledge as possible.

[4] Holley R, *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*, (National Library of Australia, 2009), http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf viewed 8 January 2010.