ARLIS/ANZ CONFERENCE PAPER, SYDNEY 2016

DATA MIGRATION LESSONS LEARNT BY THE PICTURES TEAM AT THE NATIONAL LIBRARY OF AUSTRALIA

ELEANOR GOODWIN

PICTURES COLLECTION MANAGEMENT PROJECTS OFFICER

Background

Since 1997 the National Library has been developing technical systems for the management, storage and delivery of digital material. In 2002 we began using DCM - our Digital Collections Manager. This system was purpose built for the Library's requirements and it served us faithfully for around 13 years. Technology has of course moved on and our library users now expect much more in terms of digital delivery.

With ever increasing demand for immediate access to high resolution images and features such in-depth zoom or interactive page turning – all intended to better simulate a physical object - online delivery is integral for meeting the needs of Pictures collection users.

In 2013 the National Library began developing a replacement system. The resulting new Digital Library Collection - or DLC - can cope with a greater range of file types and also will allow us to better implement modern methods of digital delivery and storage. In addition to this it offers much greater all-round flexibility, allowing it to be adapted to meet future requirements.

The project

By mid 2015 the new DLC system was ready for our Pictures material and a small team was tasked with transferring the data across from the old system. This included bibliographic data, hierarchical relationships and image metadata for more than 255,000 image files.

In preparation, data was exported from DCM into a series of gigantic Excel spreadsheets which were then carefully examined by our team for inconsistencies, oddities and errors. A detailed knowledge of the collection and our previous digital management practices was vital at this stage. An understanding of the reasoning behind certain decisions was of particular importance as it allowed us to make well-informed decisions about what to preserve and what to change during this initial planning and problem solving stage.

Where possible, we fixed whatever problems we could in advance such as top level records which lacked titles, bibliographic id numbers or other key data. Our cataloguing team was kept very busy with the updates and corrections that we requested to tidy up the records. Other issues were added to a list of post-migration clean-up tasks, either because DLC's new "Bulk edit" feature would make the corrections far more efficient, or because the required changes related to fields or functions that simply didn't exist in the old system.

During the actual migration, the new DLC system was already live and in use by the Library's printed collections and the older DCM was still being used other areas of the Library. We dealt with this challenge by placing a moratorium on Pictures cataloguing from July 31st 2015, to prevent new records or updates going into either system and causing problems.

Once the moratorium was in place, a final report was run on the data to be transferred. The report was used to sort the records into various stages and batches. We started migration with the simplest items, those with single level hierarchies or in other words, which had one image attached to one stand-alone record.

We had staff quality checking random samples in each batch as we went. This meant that we were already confident in the basic accuracy of the data transfer by the time we moved on to the more complicated hierarchies. Later quality checking could therefore focus more on checking the completeness of collections and ensuring that relationship links had correctly transferred.

The migration took a little longer than we had anticipated due to transfer difficulties with the final stage. This stage was made up of 8 oversize collections, the largest of which was the Fairfax archive which had 18,000 interconnected records.

The migration was completed on the 21st September 2015 and the cataloguing moratorium was lifted a week later once the last testing had been completed. The migration was a resounding success. During the testing and the full year that has followed we have not encountered a single example in which the data was corrupted or incorrectly transferred during migration.

In addition to the basic underlying successful transfer of data, the migration has proven to be a great opportunity to improve the quality of that data. Not only were we able to identify and correct a variety of problems prior to the migration and as targeted post-migration work, but the new system is much more transparent than the old one was. General day-to-day use of the new interface has exposed a variety of minor errors, incorrect settings and broken relationship links, which we deal with as they are encountered.

Challenges

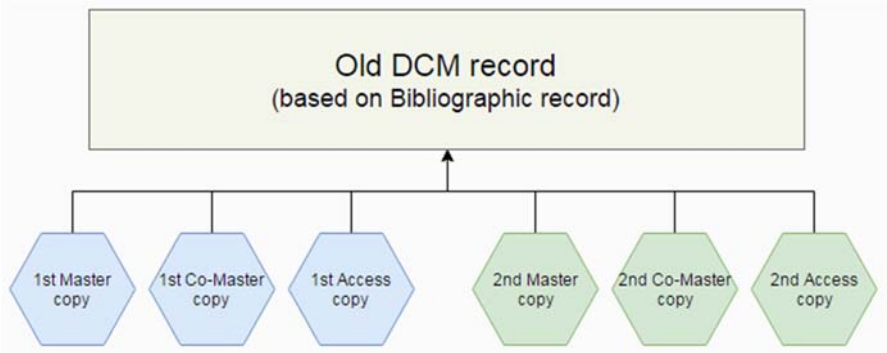
Apart from the sheer amount of data that we were attempting to transfer, our most challenging problems were nearly all caused by historic workarounds or shortcuts which forced DCM to act against its basic design.

Probably the most problematic of these was the existence of multiple master files. Logic dictates that there can only be one master file and understandably our old system was certainly designed with that in mind. Historically however, a clever way had been worked out which tricked it into accepting additional master files. This trick was rarely used, with only 62 cases across our entire

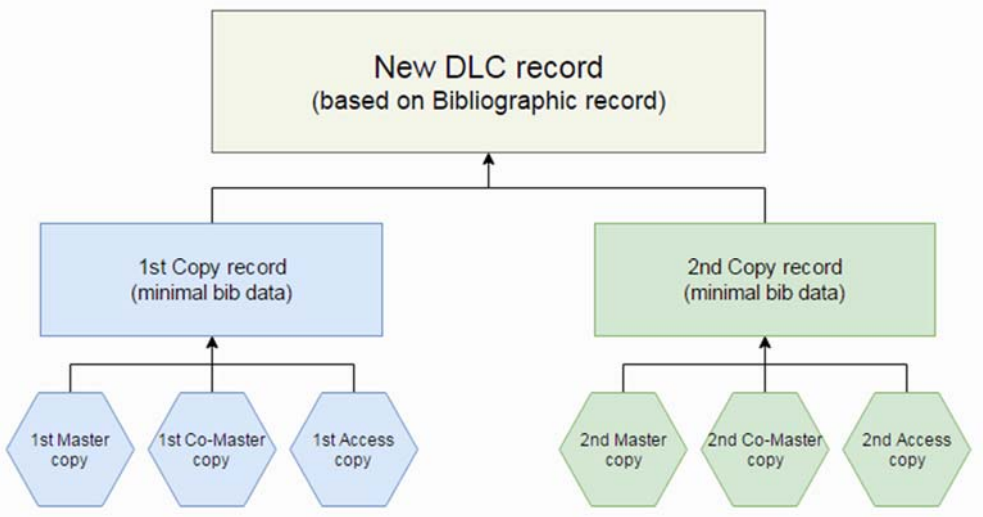
collection. They were all instances where we hold several copies of a work - often hand-coloured lithographs or photographs. Normally we would only digitise one copy as a representative of all of the others. In these particular cases however, a second or even third copy had been digitised for preservation or exhibition reasons, or due to a specific request by a researcher or publisher.

At the time, this workaround seemed like a wonderful idea as it meant we could keep all of the digital copies neatly together. Due to system constraints, accessing the additional copies was far from straightforward, but the important thing was that the files were safely stored for future need.

Of course the new DLC system was also designed to only store one master file for each record. Automatic migration for any data relating to the additional masters was not an option so instead we included only the primary copy in the migration. Then, over the course of several months, I painstakingly manually restructured the records in DLC, creating an intermediary record for each of the copies.



DCM record structure for a work with multiple digital master files



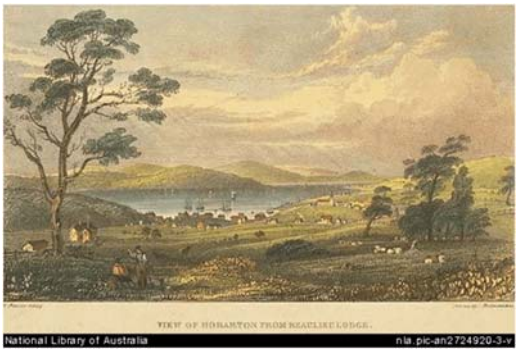
New DLC record, restructured for multiple digital master files

This new structure is entirely in keeping with how DLC works for all of our other records and interestingly enough, it could also have been used in the older system. In retrospect if we had originally chosen to set the records up this way, the data would have migrated perfectly and we would have been spared from a lot of extra post-migration work.

Surprises

The only real surprise we had from migration related to a particular set of 72MB and 18MB images. There are approximately 6000 of these from an early digitisation project and unbeknownst to any of us, they were masquerading under access copies from an even earlier digitisation attempt.

When we migrated, we did not keep the basic jpeg access copies from DCM because DLC automatically created its own access copies in the newer Jpeg2000 format. When the first of these 6000 images was encountered during quality checking we unexpectedly found that these images were displaying uncropped and in some cases with an incorrect orientation.



DCM access copy of *View of Hobarton*

DLC access copy of *View of Hobarton*

It seems that for these images, co-masters were not created and the masters themselves were somehow uploaded into DCM without triggering a regeneration of the access copy. I have no idea how this particular feat was technically achieved, but the only way to now correct the problem is to painstakingly download, individually crop and then re-upload each of these 6000 images.

Lessons learnt

All in all, the Library's Pictures migration was a very positive experience and although due credit must be given to planning and labours of our migration team, it is also clear that preparation for a smooth data migration begins long before any such project is begun or even planned by an institution. How we use our systems and the quality of the data we input on a daily basis has a significant impact on a project like this. The Library's past work practices served us well in this regard and the few oddities and exceptions in our data could be planned for as we knew about them in advance.

Corporate knowledge of your past collection management practices is vital. If you can't be assured of staff continuity (and who can?), ensure you document your oddities and exceptions; particularly, *why* something was done the way it was.

Identify and fix as much as you can in advance. If your data is consistent, it will be much easier to transfer. Poor data quality will give you a poor result in any system. It's likely that existing bad data will look even worse in a better system.

Always work within the limits your existing systems. Don't force them beyond their limits with file types, record structures, or anything else that they were not designed for. Instead look for other ways to achieve the same end result. It may take a little extra effort right now, but it could save a lot more trouble and work in the long term.

Conclusion

Through the digitisation of our unique physical works and also the ongoing acquisition of born digital material our work as art librarians is increasingly shaped by the computer systems we use to store and manage and even deliver this material. By developing and implementing DLC, the National Library has not simply upgraded a vital system, it has also better positioned itself to meet the ever increasing demands of our digital consumers both now and in the future.

It is however, still up to us as librarians to ensure that our collections are well documented and the data in our systems is accurate and consistent. This not only benefits the end users of our collections, but also helps to minimise the amount of time, effort and resources we will need to sacrifice for technological upgrades and system replacements in the future.