

Hacking the library catalogue - a voyage of discovery

Ben Kreunen
Technical Support Officer
University Digitisation Centre
The University of Melbourne Library
b.kreunen@unimelb.edu.au

Joe Arthur
Coordinator
University Digitisation Centre
The University of Melbourne Library
joseph@unimelb.edu.au

Abstract:

The University Digitisation Centre (UDC) at the University of Melbourne has been working towards making the embedding of descriptive metadata into every image and PDF a standard part of the digitisation process. This paper follows the evolution of the in-house information systems developed by UDC as we attempted to achieve this goal. Central to this has been the development of novel ways of accessing metadata from the various library catalogues, via their public interfaces. Challenges arising from the re-use of catalogue metadata in non-library systems may provide additional insights as libraries attempt to re-invent the catalogue.

Introduction

The University Digitisation Centre (UDC) at the University of Melbourne (UoM) was established from an existing imaging service (the Imaging Centre located within the Library) that provided high capacity business document scanning and microfilming services. The skills and workflows within this unit were directly transferable to the processes of digitising collection materials, making this a relatively natural progression. Early development work focussed on streamlining administrative tasks and the automation of image processing, in order to meet the challenges presented by an increase in both production capacity and the variety and complexity of items to be imaged.

Many of the major innovative projects that have changed the way libraries function (self check out, machine-sorted returns etc...) have focussed on areas where large rewards justify the large effort to achieve them, making the most of the limited availability of technical expertise. While improving data management may not have been a key driver in these situations, it has certainly played an important part in enabling effective changes to occur.

In UDC's case, the challenge of dealing with a large number of changes in a short space of time required reviewing and changing business processes without impacting on production capacity. Improving data management was a key driver for these changes, sparking an exploration of the mechanisms for collecting, processing and embedding metadata whilst new tools for managing an increasing workload were developed. The short lead-in times meant that priority always lay with streamlining workflows, but the requirements for managing workflows and metadata were so similar that the two aspects were never really separated.

Hacking is usually thought of as gaining unauthorised access to a computer system; however, in the context of this paper it is the acquisition of data from systems using non-standard methods that the system was not intended to support. Hacking has sinister connotations, but we merely wanted to obtain data for identifying the physical items in our possession and embedding into the digital outputs that we produce. Our initial goals were simply to obtain data in a format we could use, wherever and whenever we needed it.

Planning and Preparation

Many of the procedures and guidelines pertaining to digitisation have been well documented over recent years by national and other libraries and government agencies¹, but when it comes to embedding descriptive metadata into files there is still scope for improvement. Many digitised items in online repositories around the world have only a minimal amount of embedded descriptive metadata, with a large number of items having no metadata at all². As a new library service, UDC had a small window of opportunity to try different approaches to solve the metadata problem before procedures became too rigid.

The primary tools chosen for the tasks of managing data and processing files were those with which UDC had local expertise. This meant *Filemaker Pro*³ for the database engine and Phil Harvey's EXIFTool⁴ for embedding metadata. Using two whiteboards the team mapped out the process and data requirements. Digitisation workflows were mapped in a forward direction and the 'embedding metadata' workflow was mapped in reverse thus ensuring that the technical requirements for

using EXIFTool were met⁵. While it was no great surprise that the two workflows turned out to be very similar, it was still very encouraging to see that they were, at their core, virtually identical in their requirements. Keeping this learning in mind has been very important for subsequent development work. Consequently, a problem described as a workflow issue, is now often solved by redefining the problem in terms of its data management requirements and vice versa.

In the case of images, technical metadata required for the automation of embedding descriptive metadata was initially also collected by EXIFTool. The range of EXIF metadata collected was expanded to include anything that might be considered useful for other workflows (e.g. image dimensions) and was easy to collect. The value of collecting this metadata has since been realised and has been worth the small overhead of storing a few additional fields of data.

The early development stages of our database tools occurred largely in isolation from other areas of the library, and largely in ignorance of how library systems actually worked. This was partly due to most efforts being focussed on the development of new workflows and the configuration of new equipment taking priority over data management. As workflows became more established, further collaboration with other library colleagues (and other libraries) was sought enabling a gradual improvement of metadata handling. Armed with nothing more than an appreciation for the relative importance of the Bibliographic Record Identifier (BibRecID) and a username and password for the staff view of the catalogue, we produced an MS Excel template for collecting metadata of items to be scanned and a script to automatically import it into a new job record. Rather than attempt to define all of the requirements for our system at the beginning, we created a basic framework upon which to build our workflow and then embarked on a continuous cycle of refining individual sub processes. This allowed us to improve efficiency and adapt to changes as we learned more about the inner workings of an academic library. The following sections describe the various stages of the development of our in house tools and workflows.

Stage 1. Copy the librarian

UDC spreadsheet templates were distributed to collection managers in early 2010, based on our understanding that it would be easy for collection managers to locate and export records from the library catalogue as part of the process of preparing items to be sent for digitisation. The data in the initial set of completed spreadsheets were not optimal, in that they contained a number of unexpected entries. Examples included a lengthy search URL in the BibRecID field rather than the expected 8 or 9-character identifier, and titles with surplus text that had to be manually removed.

A comparison between the data in the spreadsheet with the public view of the catalogue highlighted that collection staff were not using the catalogue in the way we had anticipated. From this comparison, we surmised that many librarians:

- were unfamiliar with processes for exporting data from Millennium (Library catalogue).
- were reluctant to use the staff view of the web interface in order to see the BibRecID.
- were copying and pasting data between applications for every field, which was very time consuming.

- were accustomed to sharing spreadsheets that were human-readable but not necessarily machine-readable.

Thus, instead of saving time for collection managers time, it was taking longer than necessary to complete the spreadsheets. UDC staff then had to spend a significant amount of time cleaning up the data received before it could be imported. As an efficient means of exchanging data, this model was proving to be a complete failure.

On examination of the processes that collection managers actually used to collect the data for our spreadsheet, it became clear that it focussed on the metadata displayed in the source code of the public web interface of the catalogue (OPAC). One of the key issues was that although the BibRecID is the key identifier, it was not actually visible in the public interface. Furthermore, its presence was obscured by JavaScript in the only reference to it on the entire page. All of the data provided to UDC was nevertheless available on the web page. With the bulk of the record data being formatted within a table, it was a relatively straightforward process to set up a web portal in Filemaker and scrape the data directly from the source code of the web interface. Thus, we had at least reduced the workload on library staff preparing lists of items to be digitised.

A brief outline of what we were trying to do and a request for details on the availability of an XML⁶ feed was emailed to the catalogue administrator. Unfortunately, no usable information was forthcoming and in retrospect, the original request may have been lost in the context of the message: another dead end.

Stage 2. Getting the XML data

Minor refinements were made to the process of scraping metadata from the web interface of the catalogue but it remained a relatively crude means of collecting data that required manual clean up afterwards. In late 2011, a review of the process was undertaken to see if it could be improved. An email explaining what we were doing, with a suggestion of additional tagging of the source code to identify individual subfields, was sent to the catalogue administrator. A prompt reply was received telling us that we were doing it the hard way and that the whole process could be done much better by using the XML record.

Despite the initial frustration of having requested this information some 18 months prior, it was a very welcome nugget to add to our collection of application programming interfaces (APIs), and we were at least all on the same page. This marked the beginning of a very productive relationship with the catalogue administrator and other technical staff.

Our method of scraping the BibRecID from the source code of the public catalogue would remain in place and is still used today, but this was then used to construct a second URL to access the XML version of the record.

Construction of a separate database tool was built to deal specifically with processing the XML data before transferring the results back to our workflow database. Each XML record processed is flattened to produce an individual record of each subfield consisting of the fields RECORDKEY (BibRecID), TAG, SEQUENCE NUMBER (SN), MARCTAG, INDICATOR1 (I1), INDICATOR2 (I2), SUBFIELD INDICATOR (SI) and SUBFIELD DATA (see Figure 1).

Marc subfields							
RECORDKEY	TAG	SN	MARCTAG	I1	I2	SI	SUBFIELDDATA
b2646233	TITLE	0	245			a	Kit{u0101}b majm{u016B}{u02BB}ah ris{u0101}lah-i M
b2646233	TITLE	0	245			h	[manuscript]
b2646233	IMPRINT	0	260			c	[1831?]
b2646233	PHYS DESC	0	300			a	1 v. ;
b2646233	PHYS DESC	0	300			c	25 cm.
b2646233	NOTES	0	500			a	Insh{u0101}yi jad{u012B}d : a manual of letter writing
b2646233	NOTES	1	500			a	Holograph
b2646233	SUBJECT	0	650			a	Persian literature.
b2646233	OTHER AUTH	0	700			a	Stewart, Charles,
b2646233	OTHER AUTH	0	700			d	1764-1837.
b2646233	OTHER AUTH	1	700			a	Niz{u0323}{u0101}m{u012B} Ganjav{u012B},
b2646233	OTHER AUTH	1	700			d	1140 or 41-1202 or 3.
b2646233	OTHER AUTH	1	700			t	Makhzan al-asr{u0101}r.
b2646233	OTHER AUTH	2	700			a	Fay{243}z {u0100}b{u0101}d{u012B}, M{u012B}rz
b2646233	OTHER TITL	0	246			a	Ris{u0101}lah-i Munsh{u012B} Fay{243}zull{u0101}h
b2646233	OTHER TITL	1	246			a	Lugh{u0101}t-i h{u0323}ur{u016B}f-i tahajj{u012B}.
b2646233	OTHER TITL	2	246			a	Insh{u0101}yi jad{u012B}d.
b2646233	OTHER TITL	3	246			a	Makhzan al-asr{u0101}r.
b2646233	OTHER TITL	4	246			a	Nuskh-{u02BC}i L{u012B}{u0101}vat{u012B}.
b2646233	OTHER TITL	5	246			a	L{u012B}{u0101}vat{u012B}.
b2646233	OTHER TITL	6	246			a	Ris{u0101}lah-{u02BC}i Chah{u0101}r Gul{u1E95}
b2646233	OTHER TITL	7	246			a	Nuskha-{u02BC}i Tuzuk-i Taym{u016B}r{u012B}.

Figure 1. Sample of raw XML data after flattening
Original data: <http://cat.lib.unimelb.edu.au/xrecord=b2646233>

Getting the XML data to import into Filemaker was not as straightforward as it should have been. An Extensible Stylesheet Language Transformation (XSLT)⁷ was written to import the data directly from the URL; however, the import process causes Filemaker to crash. Checking a random XML record on W3C's⁸ XML validation service resulted in an error associated with the document type definition (DTD). This was reported back to the catalogue administrator, who passed the query on to the vendor, who replied that there were no issues with the XML structure. Irrespective of whether the vendor was correct or not, the result is the same, so an alternative workaround had to be found.

An initial proof of concept to get the import working involved a convoluted process of downloading the XML record as a text file using Wget⁹, importing the text file into a single Filemaker field, deleting the Document Type Definition (DTD) (lines 2 to 5), exporting the field to a temporary XML file and then importing this XML file. Whilst this process was successful, it was deemed to be impractical.

A second proof of concept was attempted by displaying the XML record in a Filemaker web portal and then transferring the source code of this to a text field where the data could be scraped by a script that would identify the data fields by their surrounding tags. This process failed as the University IT-supported operating

system was Windows XP, making Internet Explorer 8 the latest available version. Prior to version 10, IE automatically adds text formatting to raw XML, making the resulting source code unusable.

The use of 360 Works Scriptmaster (a free third party plugin for Filemaker) allowed the data to be downloaded from the URL directly into a text field. The use of a third-party plugin prevented us from using this process on the iOS Filemaker client. The addition of this function in Filemaker Pro 12 enabled us to explore the possibility of using iPhones as barcode scanners to create item lists in a variety of situations and locations.

Stage 3. Cleaning the data

Once again, manual inspection of the data revealed a number of challenges that had to be overcome before the data could be used. We had already decided to combine multiple subfields such as the author subfields into a single field where applicable and had calculations in place to add the required separators when necessary. MARC is, after all, designed to be machine readable, so it would make sense that similar rules would also exist in catalogues to make the displayed data human-readable.

The first obvious problem was that separators also existed within the XML data. Many years ago, our catalogue was configured to simply butt multiple subfields together without separators for display on the public interface, and generations of cataloguers have since been trained to manually enter separators into the records to make the web interface readable. A manual inspection of records of the items that we had digitised to date was conducted to derive a list of separators in use and develop a process for cleaning the data for reuse.

Character encoding was more straightforward to deal with. Diacritics are encoded as hexadecimal entities and a custom function for Filemaker was written to convert these characters into HTML entities. The command for embedding metadata into files was updated to include an instruction to treat the data to be embedded as HTML resulting in the correct character appearing in the embedded metadata.

e.g. `{u0101}` would translate to `&x0101;` and be embedded as `ā`

The display of HTML entities in our database fields made some titles difficult to read when checking item lists during digitisation. Since EXIFTool supports UTF-8¹⁰ encoded text as input data, this function was rewritten to convert the hexadecimal encoding to the actual UTF-8 characters in Filemaker. These characters remain correctly encoded throughout our workflow.

HTML entities also exist in the raw HTML entities although their occurrence is inconsistent. Ampersands for example may be represented by `&` or `&`. The frequency of these does not impact legibility of the data and so these are ignored.

Manual inspection of the data gathered for the Middle Eastern Manuscript collection led to the discovery of two invalid character encodings (“{243}” and “{246}”). The source of these errors is unclear; however, deleting them does not appear to create any problems.

The final clean-up script runs through the following steps for every subfield:

1. Trim white space from subfield data.
2. Remove trailing separators and trim additional white space.

3. Remove enclosing brackets (both round and square) and trim additional white space.
4. Remove known erroneous character encoding strings (“{243}” and “{246}”).
5. Convert hexadecimal character encoding to HTML entities or UTF-8 characters.
6. Re-combine associated subfields e.g. *AUTHOR a* and *d*.

Stage 4. Crosswalking MARC to XMP

Whilst MARC is valued for its granularity there are very few MARC fields in actual use (Smith-Yoshimura). To simplify the mapping of MARC metadata to the corresponding fields in our database (and ultimately the corresponding XMP metadata fields) we used multiple relationships built from combinations of tags, MARC tags and subfield identifiers to recombine and map the XML data to various XMP¹¹ namespaces including DC¹², IPTC¹³, PRISM¹⁴ and MWG¹⁵. This mapping is not strictly a crosswalk from MARC to XMP, as it makes use of additional tags inserted by Millennium to simplify the mapping process.

There is also a wealth of information stored in numerous notes subfields. Although mapping this data directly to other namespaces cannot be automated, for our purposes it can be used to populate lists to streamline manual data entry into additional descriptive fields.

The decisions on which fields to map to and from, and whether or not to duplicate metadata into the various namespaces supported by XMP were based on three key considerations:

1. Which fields we could actually embed into files using EXIFTool.
2. The appearance of metadata in internet search results.
3. The appearance of metadata in common applications that people will use to view the digitised files.

The requirements for these can vary somewhat from traditional library thinking, and there are many complicating factors, not least of which is the inconsistency between the interpretations of metadata between applications. To facilitate our mapping processes, we store the data in a semantic manner and calculate the actual data to be embedded at the end of the process. Imprints, for example, are split into separate fields for series, volume, issue and number, while publication dates are stored in multiple forms as month, year and complete date. This not only provides the flexibility to repurpose the metadata into additional data files for the purposes of ingesting into the various systems our clients have, but it also provides numerous ways of sorting and displaying lists of items which are useful for quality assurance and project management.

Library catalogue		UDC database	
Title (245)	The Australian musical news.	Title	The Australian musical news
Volume	v.31 ; Aug. 1940-July 1941	Volume	XXXI
		Vol. numeric	31
		Number	1
		Date	1/8/1940
		Date Year	1940
		Date Month	August

Embedded title: The Australian musical news: vol. XXXI, no. 1 (August, 1940)

Figure 2. Comparison of catalogue metadata and UDC metadata and the calculated title to be embedded into the digital files. In this instance, the files of a single bound volume of 12 issues were separated into individual issues, with additional metadata added at the time of scanning.

Typical usage considerations include:

- The appearance of multiple items with identical titles in search results is of little use to end users, so we append series, volume, issue, number and/or publication date to the end of the title when it is available (Figure 2).
- Adobe applications typically collate all embedded metadata into their corresponding field. Acrobat, for example, displays PDF:Keywords first, appending DC:Subject and IPTC:Keywords if present. To avoid duplication, only one namespace will be written.
- Available tags are selected from Dublin Core if available, followed by IPTC or PRISM.
- Many capture devices and applications also embed their own descriptive metadata. In one particular case, a scanner driver was embedding the name of the software into the artist field, which in turn was read as the author field by an application that did not read XMP metadata in PDF files. We have found it necessary to screen the metadata of all new devices for issues like this and reassign the offending metadata to a more appropriate field.

Table 1. MARC, Filemaker and XMP field mapping

XRECORD				Filemaker Field (semantic)	XMP Field	Notes
TAG	SN	MARC	SI			
TITLE	0		a b	Title	DC:Title (partial)	Actual embedded title will be followed by volume, issue, number and publication date enclosed in brackets when available
AUTHOR	all		a d	Author	DC:Creator	Semicolon-separated list. Separate calculation field checks for the existence of SI=d for each author, appending to the author name.
SUBJECT	all			Keywords	DC:Subject	Semicolon-separated list (duplicates removed). Separate calculation field appends data for SI=d to SI=a as with authors
IMPRINT	0		a	Publisher	DC:Publisher	Publisher
IMPRINT	0		b	Place of publication		Place of publication
IMPRINT	0		c	Publication date	DC:Date	Publication date
OTHER TITL	all		a		PRISM:Alternate Title	PRISM metadata is currently under consideration for use.
Other useful fields						
OTHER AUTH	all		a d			Whilst this field may be of some value to append, it is also used locally to include the name of the previous owner/source of the item
NOTES			a	Description (if applicable)	DC:Description	ERC map collection often includes a note that is appropriate for use as a description. Requires manual inspection
		856	u	URL	DC:Identifier or MWG:CollectionURI	To be decided

Stage 5. Refining and integration

Late in 2011, we were provided with another vital piece of the catalogue API: the URL format to search for a library barcode. This was a very timely addition to our arsenal, as plans were being made for the relocation of the thesis digitisation service from the Baillieu Library to UDC. We were finally able to rapidly add new items into our workflow, with a complete set of metadata automatically extracted from the library catalogue by scanning the item's barcode.

More recently, we have revisited the item record in the catalogue, following a conversation with one of our librarians about one of their digitisation projects. We had largely ignored the item record (holdings record attached to a bibliographic record) since we had started extracting data via the public interface of the catalogue. To access this data, we had to automate the login process to the staff view of the catalogue, but having done that, closer inspection of the source code of the web page showed that obtaining the two key IDs (item record ID and barcode) is possible, but not as easy as we had hoped. A subsequent enquiry to the catalogue administrator provided us with the necessary API but we are still evaluating the potential benefits this may have before considering its implementation to UDC workflows. From our experience to date, we would still have an error rate of approximately 2-5% in the list of items produced. Discrepancies between lists and the actual items listed create unwanted disruptions to productivity, and we have usually found it easier to manually collect metadata for individual items in a set as they are digitised and then deal with exceptions at the end of the process.

And so we have come full circle to find (not surprisingly) that the spreadsheets of metadata we originally sought when we established our service *can* be obtained from the catalogue.

We have progressed from requiring collection managers to complete a spreadsheet template to scanning barcodes of items just prior to them being digitised. Whilst this is helpful in reducing workloads within UDC and other areas of the library, it has also spawned some inappropriate shortcuts that may circumvent other required processes. A review of the end-to-end processes of digitisation is currently underway within the Library, and this process mapping will provide a clearer path for the management of both physical items and data throughout the digitisation process. UDC's work in improving the efficiency and quality of metadata collected will in turn provide this review with a clearer definition of the minimum data requirements at various stages, and the data management processes that we have built will provide the mechanism that will enable many of these changes to be implemented.

Case Studies:

2010: Brotherhood of St Laurence (BSL) Archives

During the first year of operation at the UDC, we participated in a collaborative project on cataloguing and digitisation of a section of the BSL Archives. Our work involved the semi-automated ingest of collections of digitised items into the BSL DSpace repository. The following aspects of the project were considered to be significant factors in achieving a successful outcome:

- The collaborative development of a cataloguing process that would meet the data requirements for all stages of the digitisation and ingest processes.

- The collection of technical metadata linked to its corresponding descriptive metadata.
- Metadata re-use methods to prepare data packages for ingesting into the DSpace repository.

Whilst the BSL project required additional customisations of our database specifically for this project, the methodology for handling data destined for 'foreign' repositories has since been applied in a reusable form to provide a consistent mechanism for dealing with custom metadata and file naming requirements. UDC now delivers files to one external and four University of Melbourne repositories using the same data processing model for each.

2012: Thesis request service

The digitisation of theses for research requests was relocated from the Baillieu Library to UDC at the beginning of 2012. The service came with a set of historical data consisting of a cardboard box with 110 CDs containing PDFs of scanned theses, a network folder of PDFs to be burnt to CD and an MS Excel list of document filenames. Previously scanned theses were physically marked with the date scanned and highlighted with a yellow marker pen. This simple practise proved to be helpful for avoiding duplication of scanning in its early days at UDC. There was a range of data issues surrounding this collection, and these issues were addressed in an unofficial project to consolidate the historical archive while streamlining the data management for the existing service. These changes included:

- copying the collection of CDs to a more secure network storage facility (approximately 5% of CDs could not be read)
- retrospectively identifying catalogue IDs for all of the PDFs
- identifying duplicate copies of theses (both digitised and native PDF files existed for some theses) and selecting a single preferred version
- merging multi-part documents to a single file set
- extracting TIFF images from preferred versions of digitised theses in the archive in preparation for reprocessing with optical character recognition (OCR), renaming and relocating to naming schema based on the BibRecID
- modifying the UDC workflow database to check for previously scanned barcodes (to be expanded to cross check for existing BibRecIDs), creating a new request by duplicating and renumbering the existing record
 - downloading the existing TIFF images in preparation for OCR if has not been performed
- incorporating AARNET's Cloudstor service via a web portal in Filemaker to facilitate direct transfer of files to clients
 - generate predefined text for each field in the Cloudstor interface to enable sending a file with a standard usage statement without having to type any data
- adding the generation of an automated email message notifying both the repository team and Special Collections of the completion of the request and the name of the file.

As with other UDC projects, the data we collect for the managing our workflows has the potential to be reused in other areas of the Library involved in upstream and downstream processes from digitisation. Some of this potential is yet to be realised but UDC attempts to promote this to other areas of the Library whenever possible.

2013: Baillieu Print Collection

The Baillieu Print Collection¹⁶, one of the University's most prized collections, was digitised in two separate projects in 2007/8 and 2012/13, the latter of which was commissioned to the UDC. A separate database was developed to prepare the metadata and cross-check prints against the 4,000 images produced in the first digitisation project. The repurposing of data collected for the project has resulted in a number of related benefits including:

- quantifying and identifying duplicate accession numbers in the existing collection prior to digitisation (an unknown number were known to exist).
- existing archival images were relocated from 4 external disk drives two floors away from the collection manager to a single network share.
- collection of region of interest coordinates for the images made archiving cropped versions of the images redundant, saving 40-50% of storage costs.
- matching of existing file names to accession number provided a searchable index of all images where previously there was none.
- methods for cropping and resizing the images for upload were reworked to streamline the processing of image requests for external clients.
- newly digitised images were available to the collection manager for distribution prior to completion of the project.
- a customised CSV export enabled the batch uploading of large batches of images including setting access restrictions automatically.
- URLs for online images were harvested automatically for quality assurance purposes.
- URLs to images and records were made available for the library communications team to share images via social media even though the KE EMU interface does not provide a display a persistent URL for records
 - metadata from cropped image sizes used to provide image URLs at sizes tailored for social media
 - URLs provided as shortened bit.ly bitmarks, using the communications team's bit.ly API key, to enable the collection of statistics from shared links.

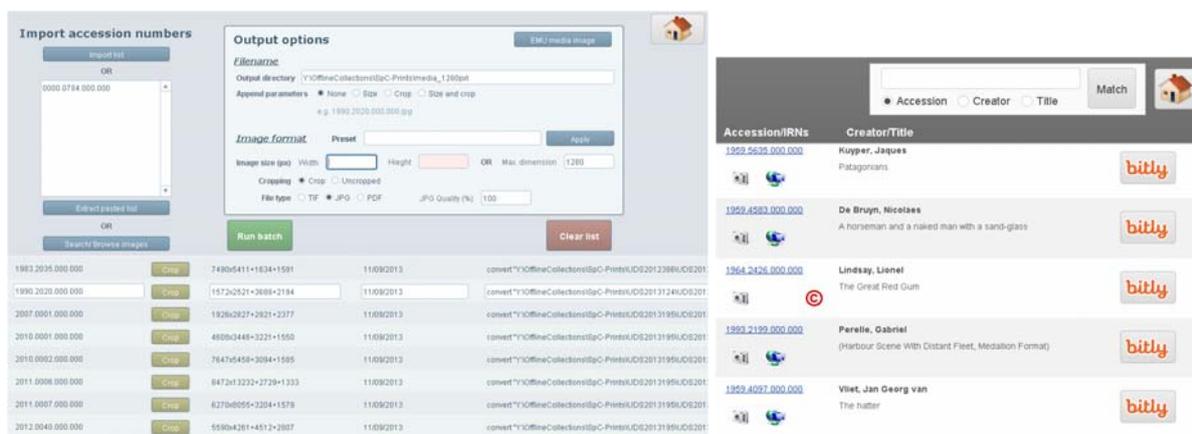


Figure 3. Screenshots showing image processing and URL sharing extensions of the original project metadata for the Baillieu Print Collection

The day-to-day management of the digitised images for this collection has been significantly improved as a direct result of enhancements to data management implemented during the digitisation process by UDC.

2010-2013: Projects managed by multiple spreadsheets

Although there has been significant take-up of database systems as data management platforms within the University Library, a considerable proportion of data management, data sharing and exchange is still based on Excel spreadsheets. The rationale for originally employing Excel spreadsheets as a metadata collection mechanism at the UDC, was partly supported by the familiarity of staff with this data format. However, the shortcomings of spreadsheets as a data management tool quickly became evident, even more so with collections which have not been fully catalogued. Some of the adversities of this approach include:

- addition of annotations in data fields to aid with manual checking can break automated updating of data, requiring further manual updating to be performed.
- keeping track of multiple versions of the same data can be extremely difficult.
- each revision cycle requires lengthy time to cross-reference spreadsheets.
- poor data quality not only slows down digitisation by making things more difficult to identify while scanning but it also increases stress and decreases motivation and job satisfaction of staff which, in turn, adversely effects productivity.

UDC is currently investigating alternative mechanisms for sharing the data in our workflow database to simplify the collection and verification of metadata. It is still too early in the process to report anything, but initial conversations with some of our key clients have been very promising.

Discussion

Over the course of the last few years, UDC has demonstrated that improvements to data management can directly translate to increased efficiency of business processes. We have also demonstrated that it is both technically possible and feasible to embed descriptive metadata into every digital file as part of any digitisation workflow if the metadata is available.

Looking back over the progress that we have made, it is obvious that it would have been a lot easier if a full set of APIs had been available for accessing data from the library catalogue from the very beginning, but the process of implementing gradual improvements over time has also provided valuable insights into our ability to adapt to changing situations and take advantage of new opportunities.

However, there are also warning signs in some of the workarounds that we have had to employ to access and reuse data from library systems. At the heart of the library catalogue is a metadata schema designed to be machine readable, facilitating the exchange of data between different systems. In order to be readable by humans, the systems presenting the data must be configured to present the data in a format suitable for reading by humans. At some time in our library's history, the system failed to deliver human readable data, but rather than correct the problem with the system, cataloguers began to change the data to compensate for the problem. The appearance of the catalogue entry on the web page may look correct but 'two

wrongs do not make a right' and this has potential impacts for every reuse of this data by other applications.

To some extent, this is indicative of many of the problems we encountered when trying to access data in less conventional ways. If the data that appears on screen can be interpreted by a person when it is accessed from the host system, then it can be very hard to convince people that there may be fundamental problems in the underlying data.

We continue to explore hacks into other UoM Library systems, including KE EMU (University Archives, Baillieu Print Collection and Grainger Museum) and Digitool (UoM Digital Repository). We have a working proof of concept where we can start with a catalogue record of a map, and by following a chain of identifiers and XML data obtain a high resolution TIFF image from our Digitool repository, even though the TIFF image is not linked to any public interfaces. In many cases, these exercises have been academic in nature but they are invaluable in exploring the boundaries of UDC's data management processes and methodologies.

Our framework of tools and methodologies lowers the effort required to find solutions to problems, and in this way we are able to apply similarly innovative thinking to some of the more day-to-day challenges we face. When exploring ways of improving processes that span multiple areas of the Library, this framework gives us the flexibility to adapt to different situations and reduce the impact of change on some of the more established practices in our library. Hacking the library catalogue and separating the data from the system has allowed us to focus on solutions to the specific challenges that we face on a day-to-day basis. Having become more aware of the possibilities provided by combining technical expertise and improved data management, UDC staff have become more aware of the types of problems that can be overcome. Rather than looking for cumbersome workarounds to get around a problem, they are now actively seeking advice to find more efficient ways to solve it. Improving data management is infectious.

References

Smith-Yoshimura, Karen, Catherine Argus, Timothy J. Dickey, Chew Chiat Naun, Lisa Rowlinson de Ortiz, and Hugh Taylor. 2010. Implications of MARC Tag Usage on Library Metadata Practices. Report produced by OCLC Research in support of the RLG Partnership.

<http://www.oclc.org/research/publications/library/2010/2010-06.pdf>

EXIFTTool user forum, <http://u88.n24.queensu.ca/exifttool/forum/index.php>
(last posted 6/1/2012)

FM Forums, Worldwide FileMaker Community, <https://fmforums.com/>
(last posted 21/7/2013)

Library of Congress Network Development and MARC Standards Office. 2008. MARC to Dublin Core Crosswalk (Unqualified)
<http://loc.gov/marc/marc2dc.html#unqualifiedlist>

Notes

¹ Commonly cited guidelines include National Library of Australia (<http://www.nla.gov.au/standards/digitisation-guidelines>) and Federal Agencies Digitization Guidelines (<http://www.digitizationguidelines.gov/>)

² A random selection of metadata samples of images viewed online:

Library of Congress (LoC): <http://bit.ly/1iHvOog>
(from <http://www.biodiversitylibrary.org/item/93835#page/7/mode/1up>)

National Library of Australia: <http://bit.ly/1lqPyw7>
(from <http://nla.gov.au/nla.map-rm317>)

National Gallery of Victoria: <http://bit.ly/19Le34V>
(from <http://www.ngv.vic.gov.au/col/work/5683>)

State Library of Victoria: <http://bit.ly/1i8l3Hw>
(from <http://www.slv.vic.gov.au/portphillip/0/0/0/doc/pp0005-001-0.shtml>)

UM Archives: <http://bit.ly/1eAFIrT>
(from <http://gallery.its.unimelb.edu.au/umblumaic/imu.php?request=search>,
and search for "UMA/I/1408")

Source files downloaded from LoC typically contain title, author, catalogue URL (as a keyword) and some administrative metadata but these are stripped out by their delivery system when images are viewed online. Other delivery systems may be doing the same thing, but either way the end result is that images viewed online often have no metadata embedded in them.

³ Filemaker Pro (<http://www.filemaker.com/>) is a relational database application for both Windows and Mac OSX. An iOS client is also available for iPhone and iPad. Whilst not formally supported by UoM IT Services, it is widely used across the campus by departments who do not have access to, or the budgets to afford database development projects.

-
- 4 EXIFTool (<http://www.sno.phy.queensu.ca/~phil/exiftool/>) is a small but powerful command line application used by UDC for reading and writing metadata in all of the image formats and PDFs that they produce.
 - 5 The technical requirements for embedding metadata with EXIFTool are relatively simple: know where the files are, know which record the files pertain to and having all of the required metadata on hand. The timing of the last requirement can vary considerably but this is not due to technical causes.
 - 6 Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.” (<http://en.wikipedia.org/wiki/XML>) As this is a text format, it is also possible to extract data using alternative pattern matching methods.
 - 7 Extensible Stylesheet Language Transformations (XSLT) is a language for transforming XML documents into another format (<http://www.w3.org/Style/XSL/>). Filemaker-specific XSLT includes instructions for importing data into specified fields in the database.
IIIRECORD to Filemaker XSLT: <http://files.digitisation.unimelb.edu.au/xsl/fm-iiirecord-import.xsl>
 - 8 The World Wide Web Consortium (W3C) is an international community that develops standards for the World Wide Web. <http://www.w3.org/>
 - 9 WGet is a small command line application used to retrieve files from the World Wide Web using HTTP and FTP <http://gnuwin32.sourceforge.net/packages/wget.htm>
 - 10 UTF-8 (Universal Character Set Transformation Format-8-bit) is an encoding that can represent every character in the Unicode character set. Not all applications support this character encoding which can result in data loss when moving data between files. Changes in the encoding of text when transferring between applications can result in data loss. The most common occurrence of this was the loss of diacritics when copying and pasting from a web browser into Excel on older versions of Windows. <http://en.wikipedia.org/wiki/UTF-8>
 - 11 Extensible Metadata Platform (XMP) is a data model developed by Adobe to store metadata properties about a file, either embedded within the file or in an accompanying file (sidecar file). The data is stored as XML and supports a number of standard metadata namespaces. <http://www.adobe.com/products/xmp/>
 - 12 Within the context of embedding metadata, Dublin Core (DC) refers to “Simple” Dublin Core, a set of 15 metadata elements comprising version 1.1 of the specification. <http://dublincore.org/documents/dces/>
 - 13 The International Press Telecommunications Council (IPTC) is the global standards body of news media. The IPTC namespace is typically used for photographs but is also useful for describing many types of files. It contains numerous elements for describing geographical locations, organisations and artworks as well as administrative and rights management metadata. <http://www.iptc.org/>
 - 14 Publishing Requirements for Industry Standard Metadata (PRISM) is a schema that has provision for a wide range of bibliographic metadata, albeit in a different

format to MARC. However, unlike DC and IPTC, few mainstream applications will display PRISM metadata.

<http://www.idealliance.org/specifications/prism-metadata-initiative>

- ¹⁵ The Metadata Working Group (MWG) is a consortium of companies, including Adobe, Apple, Microsoft, Canon, Nokia and Sony, that have joined forces in an attempt to standardise the way in which metadata is handled across different applications. A small MWG namespace was developed that includes some potentially useful fields for including references to the originating collection(s) of an item. Unfortunately, there does not appear to have been much activity from this group in recent years. Microsoft released its own photo metadata standard after joining this group. <http://www.metadataworkinggroup.com/>
- ¹⁶ Baillieu Print Collection: <http://www.lib.unimelb.edu.au/collections/special/prints/>