Opening up to the world: Moving to a web-based collection management system.

Richard Keeble, Australian Lesbian and Gay Archives (ALGA), Presented at ALMS 2019, Berlin

Abstract

There is now an expectation that collection catalogues can be searched online and content viewed from anywhere in the world, but the move to an online catalogue system is a daunting prospect for a volunteer-run community organisation like the Australian Lesbian and Gay Archives (ALGA), which has limited information technology, financial and human resources, but needs to manage 25,000 catalogued items.

What do you do when an existing catalogue started as a long-lost hand-written register, which was transferred to a card catalogue, then to an abandoned custom-built database, and finally a series of spreadsheets and is stuck firmly in the 20th century? How can existing metadata be preserved and migrated to an online software system? How should the organisation's intellectual property be kept secure? How can a software system enable volunteers with little or no information technology or library or archives experience to reliably catalogue new material? How does an LGBTI organisation share information across the world while keeping the secrets of individuals? And how can a project like this be kept running over a long period of time in a volunteer organisation?

This paper describes the practicalities of ALGA's journey from a closed, hard-to-access catalogue to a system that can be accessed world-wide and will support digital representations of ALGA's growing collection into the future.

About ALGA

ALGA is a volunteer run LGBTI+ archive established in 1978. It has a collection of over 550 shelf metres housed in over 3100 boxes, spread across three sites in Melbourne, Australia. It has over 25,000 catalogued items, with a breakdown listed below:

- 7500 books
- 2500 audio/visual recordings
- 1350 articles
- 40,000 newspaper clippings
- 350 fonds
- 1250 posters
- 700 t-shirts
- 125 artworks
- 4500+ printed photographs
- 2000 periodical titles (more than 45,000 individual issues)
- 350 theses, research and conference papers
- 3000 files of ephemera
- 300 oral histories
- 1200 badges
- 170 objects
- 100 banners

The catalogue

The catalogue began its life as a now-lost hand-written register. This was followed by a card file, which still exists and is retained as a part of the collection. In the 1990s a custom database was created by a volunteer using DB/TextWorks as a part of a university assignment, which catalogued ALGA's posters collection. However after development was completed and without the in-house skills to manage it, it was abandoned and the database moved to spreadsheets, leading to the present-day situation.

Multiple files document the collection, comprising:

- 32 Catalogue files (31Excel, 1 Word)
- 160 Finding aids (Word)
- 60 Periodicals indexes (Excel)
- Digital representations (JPG, PDF, DOC, audio, video)

The catalogue files are held on OneDrive, a Microsoft cloud storage product, which enables backup and sharing of the catalogue within ALGA, but not secure public access. Digital representations provide a partial digitisation of the collection and are held on a Windows desktop PC, and are not available online.

Excel is not practical software for an online database. To allow online access selected parts of the catalogue spreadsheets have been saved as PDFs and loaded on to ALGA's WordPress website. This allows Google to index them and make them available via Google search, but it does not enable them to be searched like a conventional catalogue using Boolean operators.

Individual spreadsheets can be searched using Excel advanced filtering, however the function is not widely understood within ALGA and only two people know how to use this functionality. Keyword searches across all of the collection spreadsheets are performed using Windows Explorer.

An increasing number of people have Excel skills, and it is easy to induct volunteers who have little previous computer experience into using Excel, providing that supervision is made available and work is monitored. However Excel's flexibility also causes problems, which were regularly observed including:

- Inadvertently leaving blank rows, breaking Excel's advanced filtering function
- The insertion of cells instead of rows which misaligns columns
- The inadvertent deletion of data, which may not be detected until after the opportunity to easily restore the data

Additionally, Excel doesn't support multi-user interaction, so only a single user can work with a collection's spreadsheet at any point in time. Excel is also impractical for storing digital representations of catalogued items.

The Project

In late 2015 academic staff at Monash University partnered with ALGA to create a collaborative community informatics project named Digital Dilemmas to research the digitisation and online accessibility of community archives, using action research methodology and ALGA as a case study[i]. The initial focus was to examine both the creation of online exhibitions and an online catalogue. At the time the volume and complexity of existing catalogue data and the lack of resources and

expertise within ALGA were perceived to make the task of moving the existing catalogue online too complex to be solved.

However, this is not a novel problem in the information technology world. A properly constituted project using existing IT project lifecycle models is suited to the task. This implies the need for IT expertise, which leads to the problem of recruiting appropriately skilled people into a volunteer organisation. This was solved with using social media. ALGA used its Facebook presence to advertise the project and this was how the author became involved at the third workshop held for the project.

The project was renamed to the more positive "Project Monte" once confidence that the project could achieve its objectives had been built. "Monte" is a reference to Monte Punshon, a much admired Australian lesbian who lived to 106 years old.

At the time of writing the project is still ongoing.

Project phases

Structured IT projects follow an identifiable cycle of activities. This data migration project followed the lifecycle below. It is iterative and does not proceed in consecutive steps because later activities will identify information that would have ideally been revealed in earlier ones. Allowing previous steps to be revisited refines steps like scope and requirements to ensure that the project produces an appropriate result in a suitable amount of time.

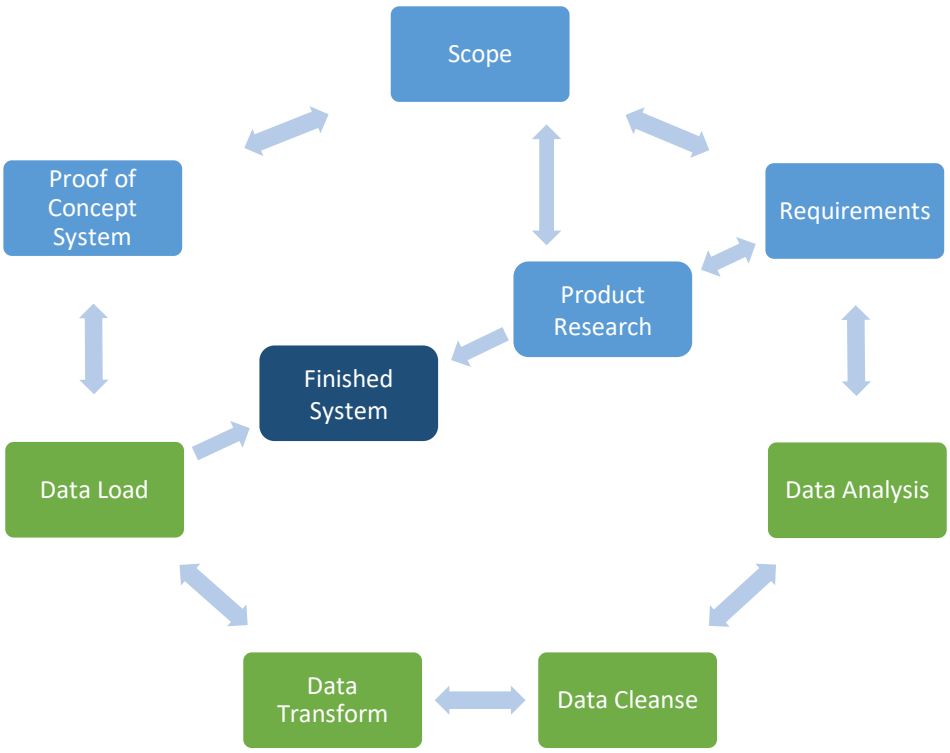

Figure 1- Project Lifecycle

Scope

The project began with workshops which identified the overall goals of the project, determined the project's constraints and identified risks to the project. Individuals performed additional work outside the workshops to gather information more efficiently than during group sessions. This work used the action-research methodology guided by the Monash University participants.

The following goals were established:

- Transfer the catalogue data to an internet accessible system
- Include existing digital representations of catalogue items, but not create new ones
- Do not integrate indexes and finding aids as database records, but treat them as attachments to items' records

Online exhibitions were excluded to be tackled at a later date. Once a catalogue is in place, online exhibitions are easier to create because item metadata and digital representations would already be available. Implementing an online exhibition system prior to a catalogue risks information captured for individual exhibitions being restricted to 'silos', instead of being easily available from a central repository for research and future exhibitions.

Constraints were straightforward, but significant. The budget was practically zero, so only volunteer labour was to be used and ideally an open source product would be selected. Regardless of product, an off the shelf system was to be used; a new system would not be created.

A risk assessment identified the following:

- Key person risk - some participants were the sole experts in the existing catalogue system. Their expertise would be unavailable if they were absent, impeding the project
- The software selected for the proof of concept might not fulfil requirements
- Development and support of the selected software system might cease
- The software hosting provider might change terms and conditions or go out of business
- The performance of hosting hardware might be inadequate
- Existing data might be lost during upload into the new system and this might not be noticed at the time
- The Archives' activities could be disrupted either because data is not available or new organisational processes needed by a new system might fail in some way
- The existing catalogue is not preserved, in the case that it might be needed for troubleshooting problems with the new system.

Scope was adjusted as the project progressed, mostly notably by excluding online exhibitions as a part of the project.

Proof of concept system - Omeka

Experiments were performed as a part of the scope process using collection management systems known to the workshop participants at the time. Three candidates were considered: Omeka, Islandora and Collective Access. Omeka was selected simply because it was easily installed on the third-party hosting servers used by ALGA for email and website hosting.

Omeka is a Web publishing platform for digital collections and online exhibitions, which was a part of its appeal when selected. It is not a proper catalogue system. It is open-source, customisable and its functionality is extensible with the use of "plug-ins", many of which are available from third-party developers. It uses Dublin Core as its default metadata schema, but can support others. It was hoped that it had sufficient functionality to function as a catalogue.

This system has been used throughout the project and provided a constant reminder of the project's feasibility and has given a sense of project progress to participants and ALGA's committee of management.

Requirements

Requirements were gathered by conducting workshops, starting with action-research directed workshops and then workshops that performed use-case analyses to identify the users of the system and the actions they performed to complete archive management tasks. As the project progressed additional requirements were observed and noted. The main requirements are listed below.

Easy data entry: Volunteers perform data entry at the Archives. These volunteers are not usually library or archives professionals and often do not stay long at the archive, so the extensive training required for a complex system is not feasible. This means that the selected system must be easy to learn and use.

Discoverability: Search functions must reliably find items in the Archives' collection and be easy to learn and use by members of the public.

Existing metadata and digital representations captured: Existing metadata includes both descriptive metadata that identifies the characteristics of an item and procedural metadata that describes things like the work required on an item and the when the record had been created or modified. Both types of data need to be accommodated. The Archives also holds a variety of item formats like periodicals, ephemera and textiles and a new system needs to capture the appropriate metadata for each type of item.

Statistical information: The archive uses statistical information derived from the catalogue to support administration and annual reporting. This information must be available directly or through access to raw data with subsequent external analysis.

Low cost / Low maintenance: ALGA has limited funds and limited access to technical support personnel. The system needs to be very inexpensive and require minimal system administration, both in initial setup and ongoing use.

Bulk data import / export: Too much metadata already existed for manual re-entry to be practical, so any replacement system needs functionality that allows this data to be automatically imported. Digital representations such as photographs, scans and A/V recordings should also be able to be imported automatically. ALGA's previous experience with its custom DB/TextWorks database created the requirement that data could be easily exported from the new system, should the system not be viable in the long term allowing - in the worst case - the data to be returned to Excel spreadsheets.

Data security: ALGA's desire is to make its collection as widely available as possible, but this is subject to a number of constraints. Many acquisitions have been donated on the condition of anonymity for the donors or embargo for a period of time. People who are the subjects of material in the Archives' may wish to not be publically associated with the material. Digital images held by

ALGA provide income, and these should be held in the catalogue system; low resolution images should be available to provide additional context to searches, but high resolution versions should not be accessible to the public. This creates a requirement that some records and/or images will be supressed for anonymous public access and a login mechanism will provide progressive levels of access.

Legal compliance: The system must provide legal compliance especially for copyright and privacy and the stored data should preferably fall under the Commonwealth of Australia's legal jurisdiction.

Data analysis

Data analysis provides an understanding of the structure and content of the data in an existing system, in this case to allow its transfer to a new system. This process began with workshops of stakeholders who had expertise in the existing system, general museum or archive skills or the IT skills required to execute the transfer. The IT specialist took the information gathered in these workshops and identified the existing structures in the data, and presented this analysis back to the project team for validation. This process also identified further requirements, which were added to the existing list.

The first task was to identify and organise all of the files that comprised the catalogue. The development of the system over time had resulted in catalogue files being mixed in with scores of other administrative files and were organised by file type (spreadsheet, text document, etc.) instead of file purpose. The catalogue files were extracted from their original locations, identified and consolidated into suitable folders.

Each of the individual files catalogued one of the sub-collections of the archives, for example periodicals, books or photographs. Once identified, the data & data structures within the files were rigorously investigated:

- The history of the metadata was sought by consulting with those at the archive involved with the creation of the catalogue files
- The purpose of, meaning of, and relationships between metadata items within and across files was established
- Similarities and differences across spreadsheets were identified
- Controlled vocabularies / name authorities were identified
- The quality of the data was assessed for completeness, gaps, accuracy and consistency

Two important ancillary files were identified: a list of ALGA's acquisitions and a list of all of the boxes housing the collection. These two files don't describe items in the collection, but provide provenance, physical location and statistical information which is essential to the ALGA's operation.

The analysis of the catalogue files indicated that the system had evolved organically without the input of formal database design expertise. Metadata structures in individual files had evolved independently of one another leading to gaps and inconsistencies in the overall structure and there had been no past comprehensive efforts to unify the metadata structures. This somewhat idiosyncratic system was highly dependent on the knowledge of the people who created it, presenting difficulties in transferring this knowledge to other people or systems.

The analysis also demonstrated the problems caused by the use of Excel instead of a purpose-built system:

- Data entered into the wrong columns
- Columns with the same meaning in different files having different metadata names
- Columns with different meaning in different files having similar names
- Controlled vocabularies / name authorities duplicated across different files, which are difficult to keep consistent
- Redundant metadata
- Metadata unlikely to be supported in an off-the-shelf system
- Spelling errors

A **data dictionary** was created from the analysis. This took two forms: a spreadsheet and a subsection of the project's website.

The spreadsheet listed every catalogue file as a row and every metadata element in each file as a column. The columns were grouped according to similarity in meaning. 781 different metadata items and approximately 35,000 records were found across all catalogue spreadsheets. These statistics gave an indication of the size of the project. This spreadsheet was invaluable in tracking the progress of the project, especially the subsequent data cleanse phase; colour coding the cells as work was completed provided a useful quick reference to determine how much work was outstanding.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	xcel spreadsheets meta data analysis						Storage Location				Identification		Title	Creator	Date	Production		Descriptive metadata						
2	Prepared by Richard Keeble, Nov 2	Record Count	Column Count	Records x Cols	Unfinished columns	Unfinished cells	Bay / Shelf	Box / Album	Page	Position on Page / Number in Box	Number	Holdings	Title	Author	Date	Place of Publication / Production	Publisher / Producer	Description	Physical format	Genre	Size			Finding aid / Index
3	Dublin Core												Title	Creator	Date		Publisher	Description	Format	Type	Extent	Identifier	Language	Table of Contents
4	Long Text																							
5	Books.xlsx	6700	29	194300	27	180900		Box Id				Copy number	Title	Author	Date	Place of publication	Publisher	Description	Format	Content ca	Pages	ISBN	Language	
6	Directories	124	12	1488	1			Box Id				Holdings	Title		Date	Place of publication	Publisher	Description	Format					
7	Theses & unpublished research p	300	13	3900	3	900		Box Id			Item number		Title	Author	Date			Description	Format	Type	Pages			
8	Periodicals	1984	32	63488	5	9920	Shelf	Box Id				Set number	Title		Date	Place of publication	Publisher	Description		Type		Identifier	Language	Index filename
9	Collections																				Extent			
10	Archival collections.xlsx	450	10	4500	10	4500		Box Id					Title	Author	Date			Description			Extent			Finding aid
11	Ephemera files.xlsx	2800	17	47600	17	47600	Shelf	Box Id					Title		Date			Description			Extent			Finding aid
12	Photographic collections.xlsx	43	7	301	7	301							Title	Author	Date			Description			Extent			Finding aid
13	Ephemera																							
14	Theatre programs.xlsx	484	13	6292				Box Id				Holdings	Title	Author	Date	Place of production		Description						
15	Calendars.xlsx	110	14	1540				Box Id				Holdings	Title	Photographer	Date	Place of publication	Publisher	Description						
16	Posters.xlsx	2520	18	45360		0	Location				Item number	Holdings	Title	Author	Date	Place of publication	Publisher	Description			Dimensions (cm)		Language	
17	Fridge magnets.xls	37	8	296	3	111		Box Id	Page No.	Number on page		Holdings	Title		Date	Place of production	Publisher	Description			Dimensions (cm)			
18	Stickers.xlsx	72	5	360	3	216		Box Id	Page No.	Position on page		Holdings	Title		Date	Place of production	Publisher	Description						
19	Postcards.xlsx	260	13	3380	2	520		Box Id	Page No.	Position on page		Holdings	Title	Author	Date	Place of publication	Publisher	Description			Dimensions (cm)			
20	Badges.xlsx	1007	18	18126	3	3021		Box Id	Page No.	Number on page		Holdings	Title	Author	Date	Place of publication	Publisher	Description			Dimensions (cm)			
21	Textiles																							
22	T-shirts.xlsx	570	21	11970	3	1710		Box Id			Item Number	Holdings	Title	Designer	Date	Place of production	Producer	Description						
23	Textiles - Banners, Clothing, Caps,	138	19	2622				Box Id			Item Number	Holdings	Title	Designer	Date	Place of production	Producer	Description	Format		Dimensions (cm)			
24	Objects																							
25	Objects.xlsx	85	6	510				Box Id			Item Number		Title	Creator	Date			Description			Dimensions (cm)			
26	Short Text																							
27	Articles.xlsx	1300	18	23400	18	23400		Box Id			Item number		Title	Author	Date	Place of publication	Publisher	Description						
28	Newspaper clippings.xlsx	9000	17	153000	17	153000							Title	Author	Date			Description						
29	Newspaper clippings - in broadsheet	276	7	1932	7	1932		Vol No.			Seq no		Title	Author	Date			Description						
30	Photographs																				Dimensions (cm)			
31	Photos.xlsx	4400	14	61600	13	57200		Album No.	Page No.	Position on page			Title	Photographer	Date	Place of production		Description	Portrait / landscape		Dimensions (cm)			
32	Audio Visual																				Duration			
33	Music Collection (published).xlsx	48	9	432				Box Id				Holdings	Title	Author	Date	Place of publication	Publisher	Description	Format					
34	Videos.xlsx	1000	21	21000	2	2000		Box Id	Tape Num	Track Number			Title	Author	Date	Place of publication	Publisher	Description	Format	Genre	Duration			
35	CDs and DVDs.xlsx	700	21	14700	21	14700		Filed in Oversize	Disk numb	Track			Title		Date	Place of publication	Publisher	Description	Format	Category	Duration			
36	Audio list.xlsx	424	18	7632	18	7632			Tape no	Side			Title		Date			Description			Duration			
37	Oral history register.xlsx [summary tab]	209	11	2299	11	2299					Interview Number				Date of interview									
38	Redundant																							
39	Scrapbooks.xlsx	25	7	175	7	175		Box Id			Scapbook number		Title	Author	Date			Description						Listed (if so file n
40	Bibliographies.xlsx	120	16	1920	1	120		Box Id				Copy number	Title	Author	Date	Place of publication	Publisher	Description	Format	Content ca	Pages	ISBN		
41	Acts of Parliament	36	13	468				Box Id				Copy number	Title	Author	Date	Place of publication	Publisher		Format	Content ca	Pages	ISBN		
42	Scripts.xlsx	36	11	396				Box Id					Title	Author	Date	Place of publication	Publisher	Description			Pages			

Figure 2- Data dictionary spreadsheet

A set of data dictionary web pages were created as a part of the project's website. As each metadata element was standardised its definition was documented describing the metadata item's purpose, format and which catalogue files used it. Standards for different data such as place name, title and author were also documented. These web pages are intended for future use to guide users in a completed system.

Data cleanse

The data analysis showed that the while the data comprehensively catalogued ALGA's collection, the data was not in a suitable state to be transferred to a new catalogue system and was not optimally organised for discoverability. A process called data cleansing[ii] is used to improve the data quality. It was structured so that the catalogue files could continue to be used as they were cleansed, so that if it was not possible to complete the project, at the very least the existing data would be in a better state, making it more accessible. This part of the project was an insufficiently recognised risk to the project schedule. The size and complexity of the data set meant that this was a painstaking, time-consuming task that became the 'biggest', and at the time of writing, still ongoing, part of the project.

Data cleanse strategies

A number of strategies were used to perform the work.

Software was written to make deterministic updates to specific columns in every row of a file. The software was a combination of Microsoft Excel formulae or formulae combined with calls to small custom Visual Basic programs.

Workshops of four to six people were held when complex, unpredictable rules across a large number of files required human interpretation. Each participant had their own computer with a subset of the catalogue files and under supervised instruction each person implemented the required changes. These workshops were named "Data Parties" under the pretext that calling a dreary task a party makes it more enjoyable, but regardless, the social aspect was of benefit because it increased participants' engagement with the project.

Where non-deterministic changes were required individuals working alone made large numbers of case-by-case changes using their own judgement. Excel's search, replace and data filtering functions were invaluable for identifying and updating problematic data for both individual work and data parties.

Case studies

Multiple categories of problem were found that needed to be resolved, each requiring different strategies. The following sections describe many of the problems found and their solutions.

Create uniform metadata names

Metadata that shared the same meaning across the entire set of catalogue files had been named differently in different files. For example, a catalogued item's title was represented using fourteen different metadata names in different catalogue spreadsheets. Another example was the use of 'author', 'designer', 'photographer', etc. in different files. In an integrated system this is likely to be represented by using a single metadata name; Dublin Core uses the name 'Creator', so these names were unified under the 'Creator' metadata name. These updates were made in data parties.

Set and apply standards

Because of the independent development of different data files, different standards had evolved for storing different types of data.

Example 1 – Author

Authors had had been represented in four different formats:

Author (Family Name, Given Name)
Tomazin, Farrah
Koziol, Michael; Massola, James

Author (Given Name, Family Name)
Nik Dimopolous

Author Family name(s)	Author Given Name(s)
Barr	James
Barreno, Horta, & da Costa	Maria Isabel, Maria Teresa & Maria Velho

Author Family Name	Author Given Name and other Authors
Lauritsen	John and David Thorstad

These formats were unified to the format of the first example. These changes were a good candidate for using a combination of Excel formulas and VB Script to reorganise the data, along with a visual check of the results in data parties to catch unusual names and where an honorific had been used.

Example 2 – Place names

Place names are commonly referred to in different ways, though these can become ambiguous, for example:

- New York
- New York City, NY
- New York USA
- New York, New York
- New York, NY
- New York, NY.
- New York, USA
- New York, NY, USA

A standard was set to represent US and Australian places as City, State, Country, with internationally recognised state abbreviations and to represent other countries place names as City, Country (no provinces). This was implemented with iterative manual inspections, using Excel's filtering and search and replace functions.

Similar creation and enforcement of standards was performed on dates, where a standard for inexact dates was implemented and for extent, where a number of standards were created depending on how an item's extent was characterised, be it a page count, dimensions, duration or shelf space.

Ensure universal metadata is in all files

Analysis identified a number of metadata items that should logically be in all files, but had been applied inconsistently. A set of 'standard' columns to be included in all files was agreed upon. The standard included descriptive metadata like Title, Creator, Date and Place of Publication as well as

administrative metadata like Box Id, Administration Comments, Date of Creation of a record and Date of Last Update of a record. These changes were created in data parties, populating the metadata where it could be sensibly extracted from existing metadata fields, or left empty, so that only new or updated records would record this metadata.

Enforce standard vocabularies

A number of standard vocabularies were identified. Some, like Subjects could be applied to any type of format in the collection, others like binding type applied only to printed material. These were standardised across all of the places where they were used using Excel's filtering and search and replace functions, and Excel's Data Validation tools were used to ensure only items from the standard vocabulary could be entered into a catalogue record.

Ensure metadata is in the correct file

Over time systems had been re-worked and the changes had not been fully completed, leaving data stranded in inappropriate catalogue files.

The fonds collection is a relatively newly developed part of ALGA's catalogue and the fonds metadata had been split between a newer, dedicated catalogue file and the Box List ancillary file which held the only record of many older fonds as sets of boxes. The data from the older fonds was extracted from Box List into the fonds catalogue file by importing it into Microsoft Access, processing it using a combination of SQL and Visual Basic and exporting it back to Excel.

The Acquisition Register ancillary file is also a relatively newly developed part of ALGA's catalogue and holds information related to the provenance of donations, including the donor's name and contact details and a unique identifier for each donation. Prior to its introduction provenance information was held in each item's catalogue record. ALGA has a preference that catalogue data does not contain information relating to donors to prevent inadvertent disclosure of that information. This part of the cleanse is incomplete at time of writing, but it is anticipated that the donor identity material will be extracted from catalogue files and added to the Acquisition Register, with newly created Acquisition identifiers replacing donor identity information in the catalogue records, ensuring that provenance can still be traced. This will most likely be done using Microsoft Access.

Consolidate related metadata into a single location and file format

ALGA's periodicals collection was largely listed in a Word document to enable publication in printed form. Additional information was contained in two other spreadsheets containing information about duplicate sets of periodicals and bound sets of periodicals.

Data in the Word document was semi-structured. Four distinct columns existed within the document, but each column contained an aggregation of metadata. Detailed analysis was required to identify each metadata element. The content of the Word document was imported into Microsoft Access and then complex text processing software was written to extract the individual elements using Visual Basic, regular expressions and heuristics. Manual review and update resolved many non-deterministic problems. The additional spreadsheets were imported into the same Microsoft Access database and further Visual Basic programs and SQL scripts were used to extract and merge this data with the main listing. The data was finally extracted to a spreadsheet.

This was by far the most difficult part of the data cleanse, because of the inherently unstructured nature of a Word document and the split of data across the different file formats.

Consolidate metadata where appropriate

Multiple metadata elements describing catalogue items had been defined inconsistently across multiple catalogue files. Information like colour, material and more generic descriptions were recorded in separate fields. This provides prompts to inexperienced volunteers to record this information and in the past allowed tightly focussed searches, but is now felt to be overly prescriptive. These metadata elements were integrated into a single Description element using Excel formulae and Visual Basic programs, with the software inserting text to ensure that the description was phrased elegantly. For example, where a t-shirt catalogue entry recorded the fields:

- material: “cotton”
- foreground colour: “white”
- background colour: “blue”

the text would be concatenated to read:

- “Cotton, white lettering on a blue background.”

Split metadata when appropriate

A number of different comments fields, usually not for public consumption, contained general administrative information, information about an item’s condition and a list of any actions required on the item. A standard set of columns were established: “Action Required”, “Condition” and “Admin Comments” and applied to all catalogue files. The generic fields in each catalogue file were then manually inspected and the metadata separated into the standard columns. This manual process was time consuming, but resulted in much higher quality data - the “Action Required” column is especially useful as it allows sets of tasks to be extracted from the catalogue files to be provided to volunteers.

Insert missing metadata

Multiple catalogue items had important metadata missing as a consequence of lax procedures. In many cases this was not vital for a successful transfer of data, with blank “Description” fields not preventing data transfer. In this case these empty fields were highlighted and volunteers set the task to progressively update the data.

In other cases important administrative information like box identification was missing. Many boxes were simply unnumbered and administrators relied on memory to locate the boxes in ALGA’s shelving system. A major undertaking was required to number all boxes and update this information in the catalogue files and in the Box List ancillary file.

Remove redundant metadata

A number of fields had been created in the catalogue files for information that wasn’t suitable for a new catalogue system. Some fields like those associated with past borrowing could simply be discarded. Other fields described the results of projects, in particular audit/stocktake fields. This type of information was agreed to be extracted into ‘past project’ files with enough information in them to identify the catalogue item and the outcome of the project. This project outcome information was not be transferred to the new system.

Spell-check

Simply performing a spelling check on metadata makes the catalogue metadata look more professional.

Data transform

A legacy system's metadata structure will typically not match a target system's metadata structure exactly, if at all. Data transformation is a process that takes suitably clean data and converts its structure to match that of a target system. It can be as simple as ensuring metadata names from the legacy system match the target system, and separate upload of controlled lists, which was the case for uploading data into the proof of concept system. Alternatively, it might require complex manipulation of the existing data to ensure that relational data is properly mapped. This task is typically performed by an IT professional with database analysis and design expertise.

Data load

Data load is a straightforward process of taking the existing cleansed and transformed data and importing it into a target system. The actual process differs depending on the types of system involved: some systems require low-level database programming to execute the transfer; other systems provide functionality to import data that has been exported from a source system into a format readable by the destination system, typically comma separated value (CSV) format.

The Omeka proof of concept system used by ALGA, provides CSV import functionality. The Excel catalogue files are saved in CSV format with column names matching the Omeka metadata schema names. These files are then opened in an Omeka import function, which detects the fields in the source file and allows an optional manual mapping of the detected metadata headings to Omeka's metadata structure. Once mapped the data is automatically imported.

Where data has been cleansed in a previous step there are almost always failures to import the entire file due to problems in the source data, usually a consequence of delimiters like quotation marks or commas in descriptive text disrupting CSV's delimiters. The existing data might also identify changes that are needed in configuration of the target system, for instance the ability to import an appropriate character set. Failed imports require either 'rolling back' the previous transfer, editing the source data files, and repeating the export-import process or selective re-import of the failed records after appropriate changes.

This task is usually performed by an IT specialist who is able to interpret and fix upload failures, although someone with generic computer operator skills can perform the work when data can be reliably uploaded without error.

The iterative process of loading, fixing and reloading data provides partial quality assurance of the data cleanse and load processes. Another check is required to ensure that all records and each metadata element in the records have been uploaded. Ideally this is performed by exporting the data from the new system in a format that can be matched with the old system, and using file comparison software. At the very least it requires comparison of record counts in the original and target systems and manual spot-checks of records to ensure no gross errors have occurred during data transfer.

Once a target system is selected and the data is fully cleansed a final data load is performed prior to the new system 'going live'.

Product research

Product research identifies and assesses available products to determine how well they fulfil system requirements. Product research leads to realistic scope and requirements. If gaps are found between a product's functionality and system scope and requirements then compromises are needed either in the choice of software system or expectations of what the end system can achieve. A small organisation must accept the products on offer and adjust operations to suit the product, not the other way around.

ALGA has been recommended products by word of mouth and targeted internet searches have identified others. The first assessment of these products has been made from information on the providers' websites. Many products are available as online demonstration systems and these have been used to manually enter realistic data from ALGA's catalogue and, when the system supports it, small uploads of data have been attempted. For closed-source proprietary systems where online demonstration systems are not available, meetings with vendors have expanded on the information available from their websites and allowed the vendors to demonstrate the systems. Where products have been found to merit a more detailed examination and where practical, demonstration systems have been installed, which assesses both ease of installation and maintenance, and day-to-day functionality.

More than a superficial assessment of available systems is required. What looks good on paper might not fulfil subjective requirements such as ease of use. Word of mouth suggestions - made with the best of intentions - might not be made with sufficient knowledge of the products or organisation implementing them. Many off-the shelf systems are targeted at a single type of collection format such as books or fonds, and will not work well with all the gallery, library, archive and museum formats that ALGA holds.

Some products can be customised to better suit an organisation's requirements. This is risky, because modifications can be disrupted or even lost when the base product is updated to newer versions, so modifications must be carefully managed. If an organisation doesn't have the resources to manage customisations, it may be better to avoid them and accept a less functional product. Third-party extensions (sometimes called plug-ins) are essentially customisations made available to a wide audience and often distributed using the same channels as the main product. These also come with management risk as they may also not function after an upgrade to the main product and might become unsupported, though conversely in open source software they may be adopted as a part of the standard installation.

As well as the product, the products' native meta-data schemas require investigation to ensure that existing metadata can be supported by a new system. Vendors often use a proprietary schema which requires specialised data transformation, a task that might require payment to the vendor. Proprietary schemas also risk vendor lock-in, making it difficult to move away from a no-longer optimal product. Two metadata schemas have been notable in ALGA's research. MARC is a schema designed for library bibliographic metadata. It was originally developed in the 1960s and retains many of the technical constraints of its era, making it difficult to work with compared to newer schemas. Dublin Core, developed in the mid-1990s, is a simpler metadata schema designed to represent both digital and physical resources and exists as a simple set of core metadata elements and an extended set. It does not allow for procedural metadata, requiring customisation for use in a collection management system. Its online documentation is mostly a statement of the standards and there are few practical examples of its use, especially for the extended version. This can make the standard difficult to understand for a novice.

At the time of writing, product research is still ongoing – a product has not yet been selected, but the following products have been investigated.

- Omeka
- Collective Access
- eHive / Vernon
- Archives Space
- Islandora
- Access to Memory (ATOM)
- Inmagic
- Recollect
- Library Thing
- Issuu
- Koha
- AndOrNot
- Occams
- Victorian Collections
- kEMU
- Archivematica

Omeka has been the most investigated product, as a consequence of its selection as a proof of concept system and its weaknesses became apparent through use, notably security and the ability to store all of ALGA's existing metadata. Omeka has security limitations, in particular the ability to hold, but restrict high-resolution digital images, and the ability to restrict entire records as being embargoed. Omeka's use of Dublin Core highlighted the inability to accommodate the procedural data required for collection management, such as storage locations and administrative activity, however plug-ins make up for some of the shortfall by providing information such as record creation and update history. Customisations were made to successfully allow additional metadata to be recorded, but the risks of these failing after an update questions their suitability. No other large collections could be found that were using Omeka as a catalogue system, however some small organisations were identified that were making effective use of Omeka for digital exhibitions.

Human factors

This project has produced tangible results on a very low budget, but it has required many hours of carefully organised volunteer work, both in groups and individual efforts. The group work has primarily set the direction of the project and captured organisational knowledge like policies, procedures and systems. This work has taken place on weekends. This has been guided by suitably skilled people to ensure that group sessions stay focussed on the task at hand and that the required information is captured. Independent work has been performed for tasks unsuitable for efficient collective effort like data analysis, software development and hands-on data cleanse work, however workshops where groups have worked on guided individual assignments made work more enjoyable and gave a sense of camaraderie.

Volunteer engagement has been variable. Some volunteers attended only a small number of workshops and were observers rather than participants in the project. Others have attended when called upon for specific skills, but others, mostly with a long association with ALGA, have formed a core group which has been essential to the progress of the project.

Different strategies have been used for communication between volunteers, with varying success.

- Email
- OneDrive: File storage and sharing
- WordPress: Workshop reports, research reports, reference sharing
- Zotero: Group discussions, reference sharing, document collaboration
- Slack: Instant messaging
- Doodle: Meeting scheduling
- Facebook: Social media

There was overlap between WordPress and Zotero. They were useful for enabling easy access to documents, but without dedicated training, hard to use for volunteers. Slack was advocated, but some volunteers had difficulty installing it and because the group did not need to be in constant communication, it use was quickly abandoned. Doodle was partially successful, but requires user engagement to work well. The most used and essential mechanisms have been Facebook, email and OneDrive. Facebook enabled the recruitment of key volunteers. The universal availability of email meant that participants required no training nor needed to install software on their devices to communicate with each other. OneDrive was already well established at ALGA and key participants already knew how to use it.

Within the core group of volunteers there have been interruptions to participation because of events in the participants' lives. An early participant relocated to another city for work, ending her association with the project. Another took sabbatical leave abroad. Two key participants experienced bereavements, making them unavailable for extended periods of time. Another interruption took a different form: the not for profit organisation that hosted ALGA's reading room and storage space sold its building, necessitating ALGA's relocation. The author became one of two relocation managers, which stopped work on the project for more than six months.

These interruptions have created at least twelve months of delay to the project, without which the project would likely be complete. When these events occur, the ability to restart and rebuild momentum is important. This ability was established early in the project by creating and maintaining stakeholder and participant engagement. IT expertise available at the start of the project created a project plan of small, tangible steps each having observable results. This convinced stakeholders of the project's potential success, controlled the scope of the project, and ensured that participants weren't overwhelmed by the overall complexity of the project. The step-wise nature of the project meant that it was easy to see where to resume work after delays.

Conclusion

ALGA's experience moving an offline, file-based catalogue system to an online system that captures digital representations of catalogued items has proceeded with little financial outlay, but has required a core group of volunteers who are prepared to dedicate large amounts of time over many years. Patience and flexibility have been required to accommodate events in the lives of participants that have significantly delayed the project, however the dedication of the core team has ensured that the project has continued. The recruitment of volunteers with information technology project skills introduced a formal project lifecycle, enabled rigorous analysis of the existing system and rigorous evaluation of potential replacements. This recruitment was enabled via an already active online social media presence.

The state of ALGA's existing data has been a significant unrecognised risk. The data has needed careful scrutiny to evaluate how ready it was to be transferred to a new system and the necessary data cleanse using both software and manual intervention has required large amounts of time. As an output of data analysis, a data dictionary has proved invaluable in both understanding existing data and managing the data cleanse process.

The use of a proof of concept system has allowed validation of scope, requirements and the condition of the data to be transferred. Rigorous product research has affected both the scope and stated requirements of the end system, and has avoided the selection of inappropriate products.

Other organisation's projects may differ in scope, requirements or data volumes, however the approach taken by ALGA can form a basis for other data migration projects, where existing metadata is to be retained.

---

[i] Ruge, C. et. al. (2016). Digital Dilemmas: a participatory investigation into developing a digital strategy for a community archive. Retrieved 12/07/2019 from https://www.vala.org.au/vala2016-proceedings/vala2016-session-13-ruge

[ii] https://en.wikipedia.org/wiki/Data_cleansing