Technical Report: Doctoral Theses Digitisation

Ingrid Mason, Digital Research Repository Coordinator
New Zealand Electronic Text Centre, University Library
Victoria University of Wellington, New Zealand
September 2008

Introduction

A doctoral theses digitisation project began in March 2008 and was estimated to be complete by November 2008. The phases (some overlapping) of the project were:

- contract settlement
- database development
- batch processing (itemisation and transportation)
- letter generation (permissions)
- item processing (receipt of digital files)

The planning, implementation, business and technical details of the thesis digitisation project are described in this report.

Introduction 1
Project summary 1
Project proposal 2
Project planning 3
Project scope 3
Digitisation database........ 4
Project issues........ 4
Broader issues 5
Project tools 6
Project findings........ 6
Data gathered 7
Appendices 8

Project summary

Doctoral theses (~1200) in the University Library's collection have been digitised and uploaded into the Library's two research repositories: RestrictedArchive@Victoria and ResearchArchive@Victoria. With a view to sharing learning and useful information key considerations for other tertiary institutions undertaking a similar project are:

- digital file sizes and server storage space
- purpose of and standards of digitisation for access
- data matching from library system and alumni database
- database listing and tracking of theses and allied tasks
- inventory listing and batching of theses into boxes
- costs for digitisation, transportation and short term assistance

Digital file sizes on average are: ~35MB. There are two 'modes' of file size ~20MB and ~30MB. A direct proportional correlation between page numbers and file size was expected and is evident. The scanning levels are 300dpi for theses with good scan readability and 400dpi where readability is marginalised by the quality of the text. Standards for quality assurance were 100% metadata accuracy and 98% OCR accuracy for 98% of the works digitised.

All items were inventoried in a separate database to record all movement, actions and tasks associated with the digitisation of the theses and the parallel permissions request (to make the thesis openly accessible). Data was drawn from the library system and matched against contact information drawn from the alumni database to maintain a record of what theses could be copied from the restricted access repository to the open access repository. Allied tasks relating to the movement and return of theses to the Library and hyperlinks added to the library catalogue record were recorded in this database to track workflow and for reporting.

Only in theses (~pre 1984) have issues of character recognition arisen. Batches of ~8 theses per box were sent to the contractor. This number of theses per batch was determined by box size and the ability for a strong person to lift and move a box of ~25kg for transportation purposes. Both courier and bulk methods of transportation have been employed to ship the theses to and from the contractor.

Digitisation costs per page were NZ$0.11 per page exclusive of GST. Transportation costs varied depending on the level of service. The bulk shipment of ~100 boxes of theses within New Zealand was ~NZ$420. The courier shipment costs ~$35 per separate box for subsequent batches.

Project proposal

The project to provide digital access[1] to the doctoral theses in the University Library's collection was proposed to enable wider access, in digital format, to the theses within the University, and where feasible, outside the university. The benefits accruing from this being:

- digital access to researchers via the university network or the internet to a digital version[2] of the thesis
- maintenance of a print copy of a thesis for preservation purposes and where access to the print copy is appropriate or preferable
- promotion of the University's post-graduate research via the open access research repository ResearchArchive@Victoria[3]; research hubs such as the Kiwi Research Information Service[4] (national) and OpenDOAR[5] (international); and search engines such as Google[6], and in particular, Google Scholar[7]

[1] Digital access means either via the university network to either restricted or open access research repositories or via the internet to the open access research repository. The university library implemented two instances of DSpace to manage the library's digital collection items for internal and external use: RestrictedArchive@Victoria and ResearchArchive@Victoria (http://researcharchive.vuw.ac.nz) respectively.

[2] Digital version means either a facsimile of the thesis through the digitisation of the print copy or the acquisition of a digital copy of the thesis in its original MIME type (and the conversion of these digital files into PDF) or the acquisition of the thesis in PDF. There have been issues with the acquisition of theses in PDF where the files are not readily indexed by search engines such as Google, see: http://www.rsp.ac.uk/repos/tools Last Accessed 03/09/2008

[3] http://researcharchive.vuw.ac.nz Last accessed: 29/08/2008

[4] http://www.nzresearch.org.nz Last accessed: 29/08/2008

[5] http://www.opendoar.org/ Last accessed: 29/08/2008

[6] http://www.google.co.nz/ Last accessed: 29/08/2008

[7] http://scholar.google.com Last accessed: 29/08/2008

Project planning

Liaison with colleagues at University of Auckland over their experience with digitising their doctoral theses provided invaluable insights into efficient methods and issues that may arise in the project. This openness enabled replication of some aspects of their approach and deviation where required to meet the specific requirements of Victoria University of Wellington or to avoid known pitfalls.

This project required internal and external liaison. Internal liaison and consultation was with Information and Technology Services (technical environment), the Library (business and technical management) and the Alumni Office (permissions process). All areas of the Library were liaised with to enable smooth integration into business and technical processes and to draw upon the expertise, authority and time of colleagues. External liaison was with a contractor commissioned to undertake the digitisation of the theses and produce a digital file and metadata ready for ingest into the library's research repository and process ~150 theses per week.

Project scope

The scope[8] of the project was determined by the budget, the number of theses in the collection and the cost of each thesis digitised:

- the capital budget allocated for the project was NZ$50,000
- some operational costs for professional and assistant time and transportation costs were absorbed in other appropriate Library and University budget allocations
- the estimates for the cost of digitising each thesis[9] was NZ$30 (based on average page count)
- the number of doctoral theses in the thesis collection is ~1200
- the large format (A3+) and supplementary material on disks have not been digitised or and copied into the thesis digital file[10]
- digital files to be available via FTP server from contractor for download

The tasks undertaken by University staff:

- issue and return of theses by Campus Library and Lending and Reader Services librarians and library assistants
- movement of theses within the university by Campus Care
- batching up of theses into boxes and permissions letter and email generation by part-time library assistant
- transportation of theses by administrators
- batch uploading into DSpace by system administrator
- copying of theses (with open access permission) into ResearchArchive by Catalogue Librarians

[8] Theses with large format or supplementary materials, i.e. fold out maps or discs were out of scope due to cost (~NZ$100 per item with A3+ print material) and issues with file formats. The opportunity to include this material with the digital file of the text of the thesis at a later date is under consideration. This issue arises in the main with theses developed in the sciences, in particular geo-science.

[9] There are page limits set for theses by the library which constrains page numbers. Some theses though are multi-volume and include appendices or supplementary material.

[10] A special project will need to be undertaken to manage the digitisation and copying of this material into the thesis file. Large format content may require different scanning specifications and maintained as separate digital files associated with the thesis digital file. The digital content on disks will require appropriate applications to render and make accessible.

- rebinding of theses by Collection Services staff
- project management by Digital Research Repository Coordinator and project lead Director, NZETC

Digitisation database

A database was developed to manage data imported from the library collection system and the alumni database. This workflow and inventory database recorded the batching up and movement of theses to and from the contractor (back into library storage); receipt of digitised files of theses from the contractor; permissions to make the theses open access sent and received from alumni; ingest and archival into the repository; and electronic links added to the library catalogue record. See data structure in the appendices.

Project issues

Coincident projects

At the same time the digitisation project was being undertaken the library was assessing its physical collections and moving low use collection material to offsite storage. The second copies of the theses (used for digitisation) were housed in the same collection area that part of the collection assessment project was addressing at the central campus library. The coincident timing of these projects meant that staff availability for assistance was low and flexibility in approach and task allocation was required to ensure that both project managers liaised and coordinated their planning, physical handling, access to the storage area, circulation control on the library system and project timelines.

Diverse storage areas

Aside a closed stack collection at central (with second copies of most of the doctoral theses), theses (both copies) were also housed at three of the other campus libraries, a special library on central campus, and some theses had only one copy and were housed in the reserves collection at central. The batching up and retrieval of these theses needed to be carefully managed with the campus library staff to ensure it fitted in with their business planning, staff availability and provision of user access. The digitisation database developed enabled inventory control and provided a means to group series of theses, track their movement and return, and enable efficient interoperation with the library staff involved with the physical management of and user access to doctoral theses in diverse collection storage areas.

Unique material

The library has two copies for 98% of its doctoral theses. Where the library has only one copy of a doctoral thesis it was important to ensure that these were processed and returned by the contractor swiftly. These unique copies were sent up in a special ‘job lot’ by courier to the contractor and processed within one month. By comparison, the rest of the second copies are being processed systematically (approximately 150 per week). This urgency for the unique copies was with a view to reducing constraints on access and to ensure that there was no loss of unique material held by the library. This issue of potential loss of unique material will arise for the library in the proposed digitisation of the master’s thesis (where the library on retains one copy of the thesis). To avoid risk of loss there is a clear need for careful inventorying and tracking of unique works sent offsite (for digitisation) and returned.

Availability form

The library has a policy of loading the availability form (which includes permission to make works openly accessible) in digital form into the repository along with the digital file of the thesis. Access to this form is restricted to system administrators and is not publicly available. During the digitisation of the theses it became apparent that the contractor was scanning the inside covers of the thesis. A request was made to the contractor to begin scanning from the title page and to avoid scanning the form. Where a thesis had (inadvertently) had the availability form scanned in the digital file, this page was removed by cataloguers when the digital file of the thesis was copied into the open access repository (as permission to make it openly accessible had been gained from the author).

Supplementary material

Some theses include supplementary material, e.g. tapes, CDs or foldouts. Most commonly the sciences, in particular, the geosciences large format maps form part of the thesis. The scope for the project did not include the payment for scanning large format (A3+) material or to shift content off portable media. During the process of digitisation, where supplementary material forms part of the thesis, this has been recorded. This recording is done with a view to budgeting for further work to be done to process this supplementary material and include it with the rest of the thesis and affects ~8% of theses.

Digital files

In the letter to request permission from alumni, a request was made for a digital file (if one exists of the thesis) to be supplied. The benefit of this is that the file size is often smaller (for the download benefit of users). Volunteers often supply several files which need to be compiled in the correct order, which takes time. Sometimes the entire thesis is not available in digital form which means that the offer to be supplied digital files needs to be rejected. Sometimes the files supplied are in PDF or LaTex or PostScript files that are not possible to index by search engines, either in rendered text or in fact encrypted or restricted in some manner, which means these files have to be discarded or open PDF files requested. A better approach would have been to supply a guide on what digital files were best offered in with the permission letter to the alumni to avoid confusion.

Broader issues

- Managing change in the mode of access and impact on print and electronic resources and service support
- File formats and digital asset management (access and preservation)
- Coordination of multiple projects in library planning and management

Project tools

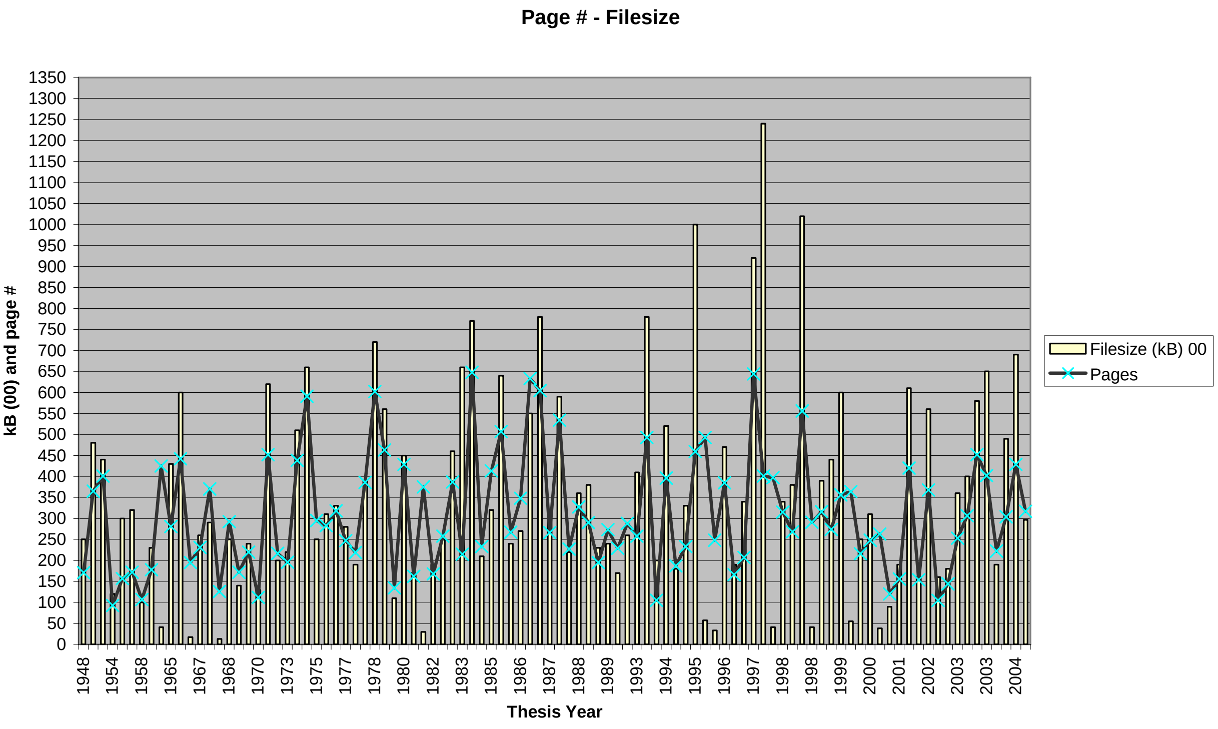
- The digitisation (workflow and inventory) database was constructed with Microsoft Access 2003.
- The digital files supplied by alumni (where required) were converted to PDF format with Adobe Acrobat 7.0 Standard.
- The scanning resolution was checked (in the quality assurance test) with Adobe Acrobat 6.0 Professional.
- The file transfer from the contractor was undertaken with FTP-32 (client for Windows)[11] and FileZilla[12].
- The zipped files supplied by the contract were unzipped (and files extracted to separate folders) with 7-zip[13].
- The repository application storing and providing access to all of the digitised theses is DSpace[14].

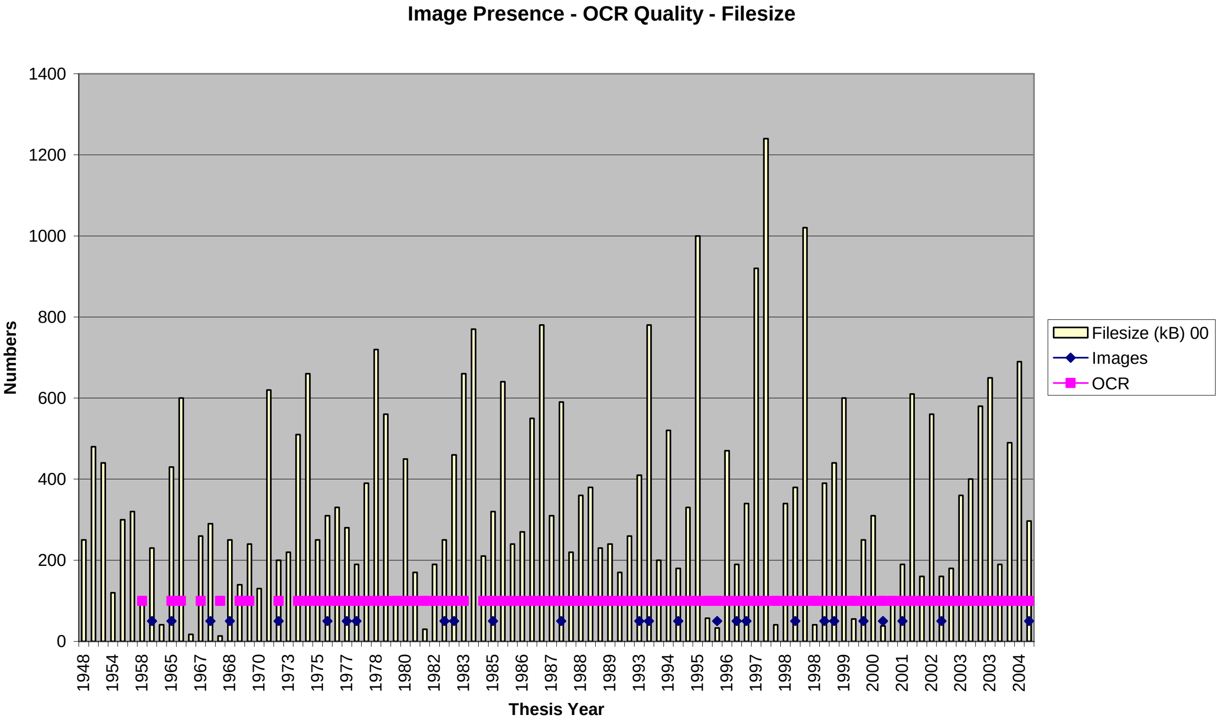
Project findings

- On average digitised theses are ~35MB in file size so file transfer and storage can be calculated.
- Images have been included in theses regularly over the time period doctoral theses have been produced. Expectations that they would increase due to greater availability of tools to include them have not been realised.
- OCR and legibility has been consistently maintained at a reasonable level (using the increased dpi levels where required) from the mid/late 1970s onwards. Expectations that OCR and legibility would decrease with theses typed (prior to computer age and printing) have not been realised. Though less success has been had in scanning and OCR with theses developed prior to 1975 for legibility and fulltext searching.
- On average theses are ~309 page numbers, so if pricing is calculated per page, this can help with estimates.
- Developing the digitisation database and maintaining data relating to the workflow ensured that no theses were lost and all theses could be tracked during the project.
- Early circulation of a clear proposal and plan ensures the ability to work in with colleagues that already have other priorities and projects to manage.
- Requests for interlibrary loan for the print version from the university library have slightly decreased, this is attributed to having more theses openly accessible via ResearchArchive@Victoria for searchers to retrieve directly themselves.

[11] http://www.gabn.net/junodj/ws_ftp32.htm Last accessed: 03/09/2008
[12] http://filezilla-project.org/ Last accessed: 03/09/2008
[13] http://www.7-zip.org/ Last accessed: 03/09/2008
[14] http://www.dspace.org/ Last accessed: 03/09/2008

Data gathered





Appendices

1. Quality Assurance Test
2. Workflow diagrams
3. Database structure

Quality Assurance Test

Work standards

- 300 dpi images/illustrations/text
- greyscale and colour
- poor quality text 400+ dpi
- 100% metadata accuracy
- 98% OCR accuracy

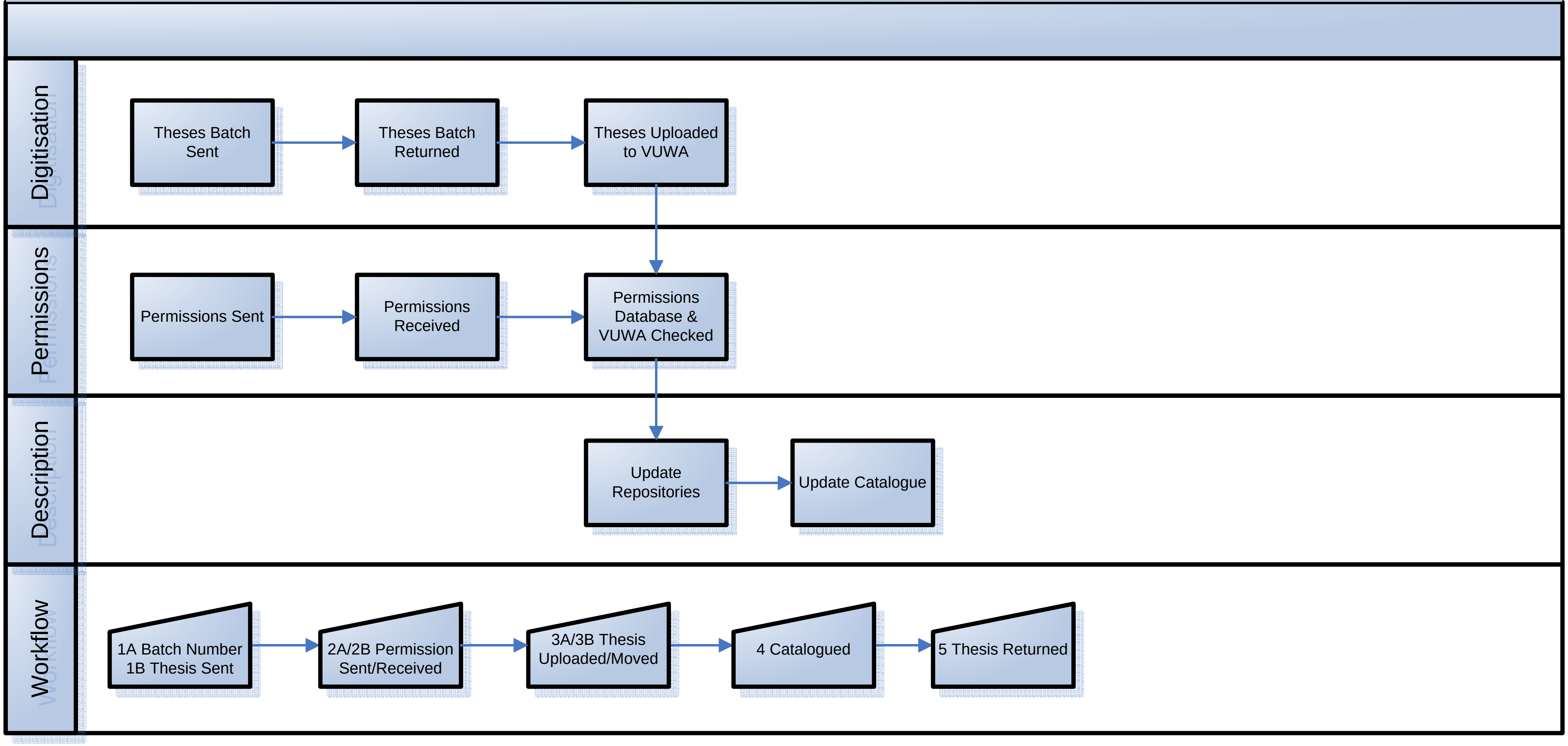
Test

- 10% of every batch is tested randomly
- open all three files: content, XML and PDF
- content file should contain 'thesis.pdf'
- record total page numbers
- record file size
- record year of thesis
- OCR – test 3 words that appear in the front (abstract), middle (page 150) and end (bibliography) of the document, pay attention to typescript works, or text that is blurry
- scroll quickly through to check images and their quality
- compare metadata with that in thesis – pay attention to symbols and characters

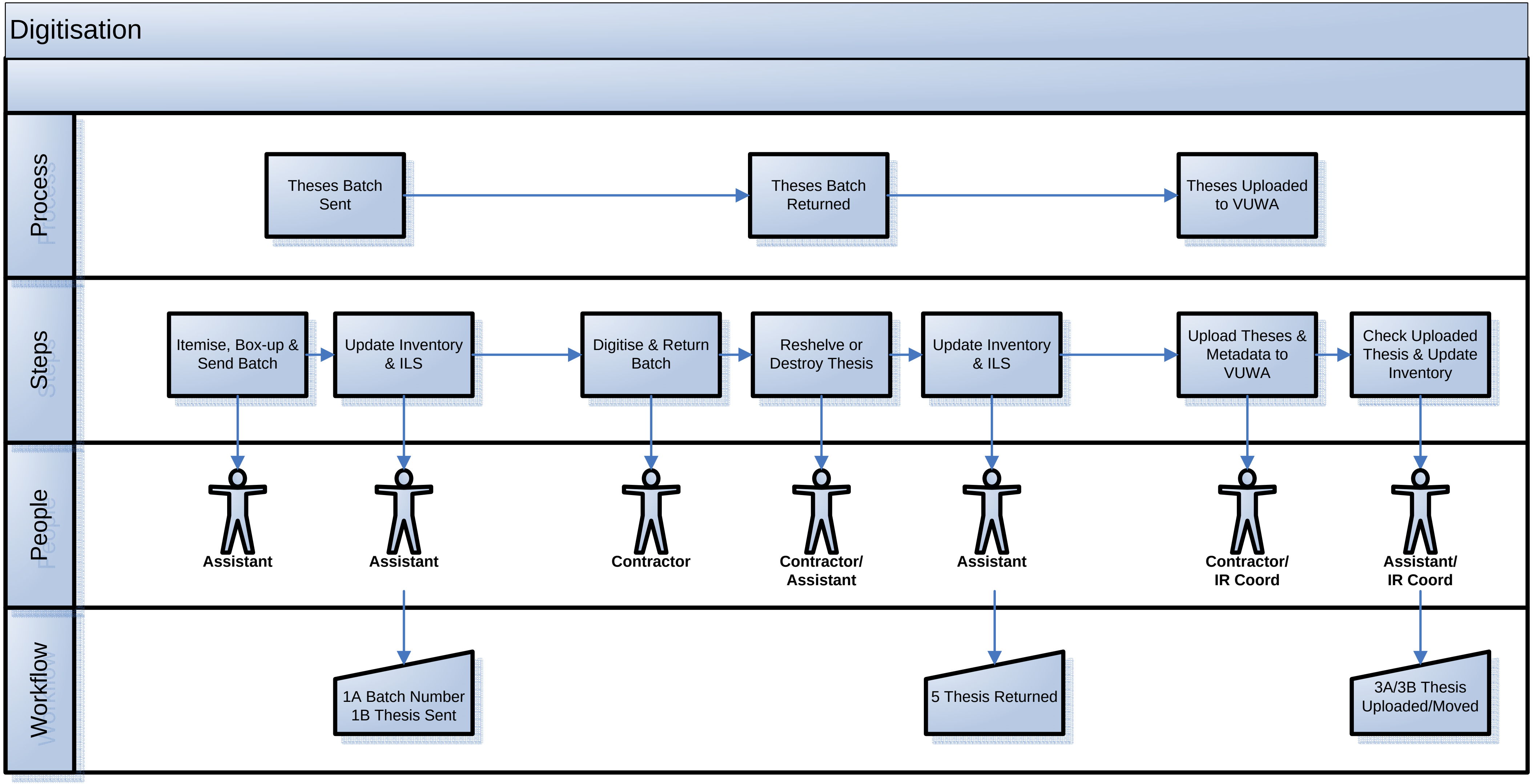
Check

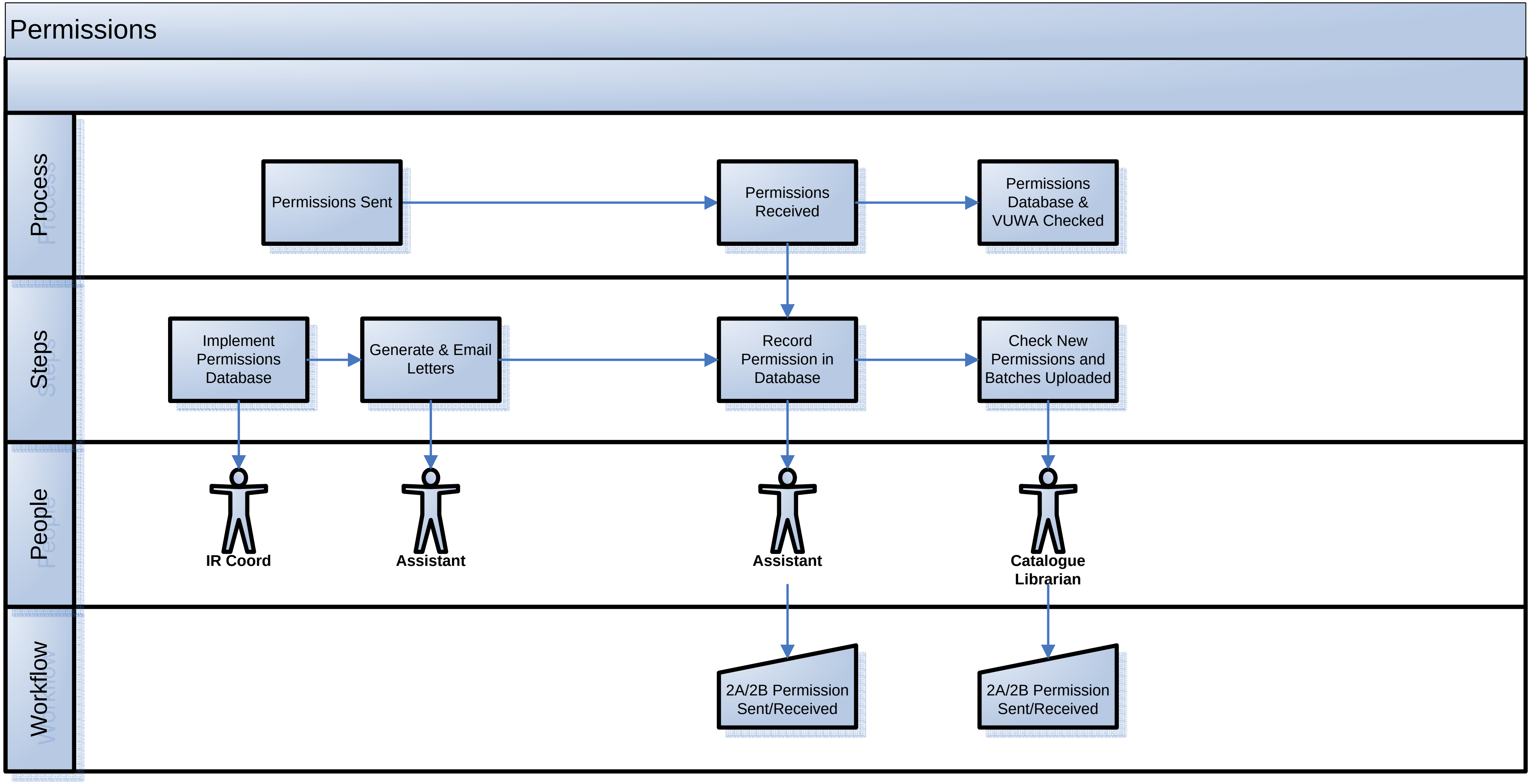
- work log from contractor for scanner's notes and check with multi volumes, that the first XML file has the abstract (not the second XML file)

Thesis Digitisation - Overall

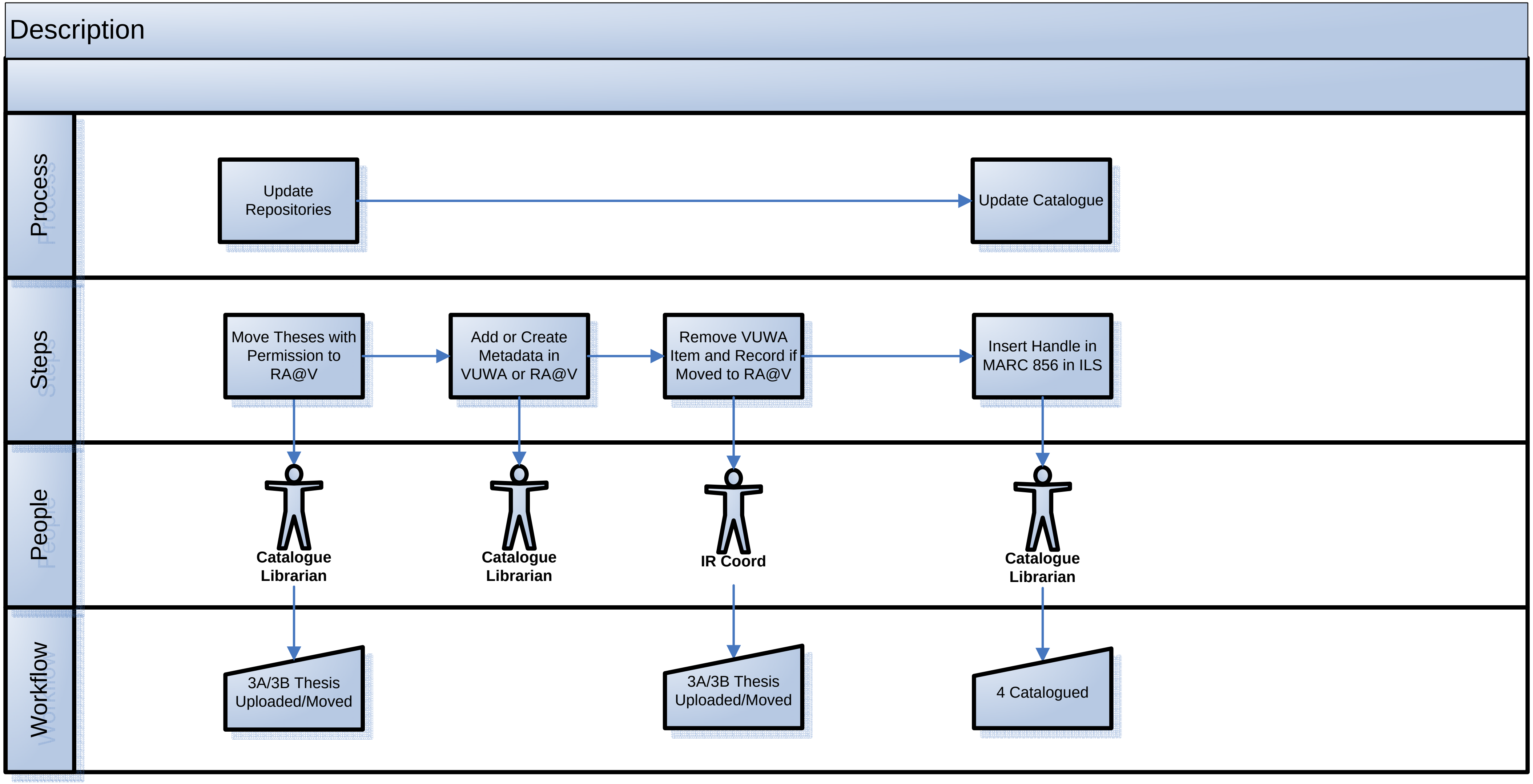


Digitisation

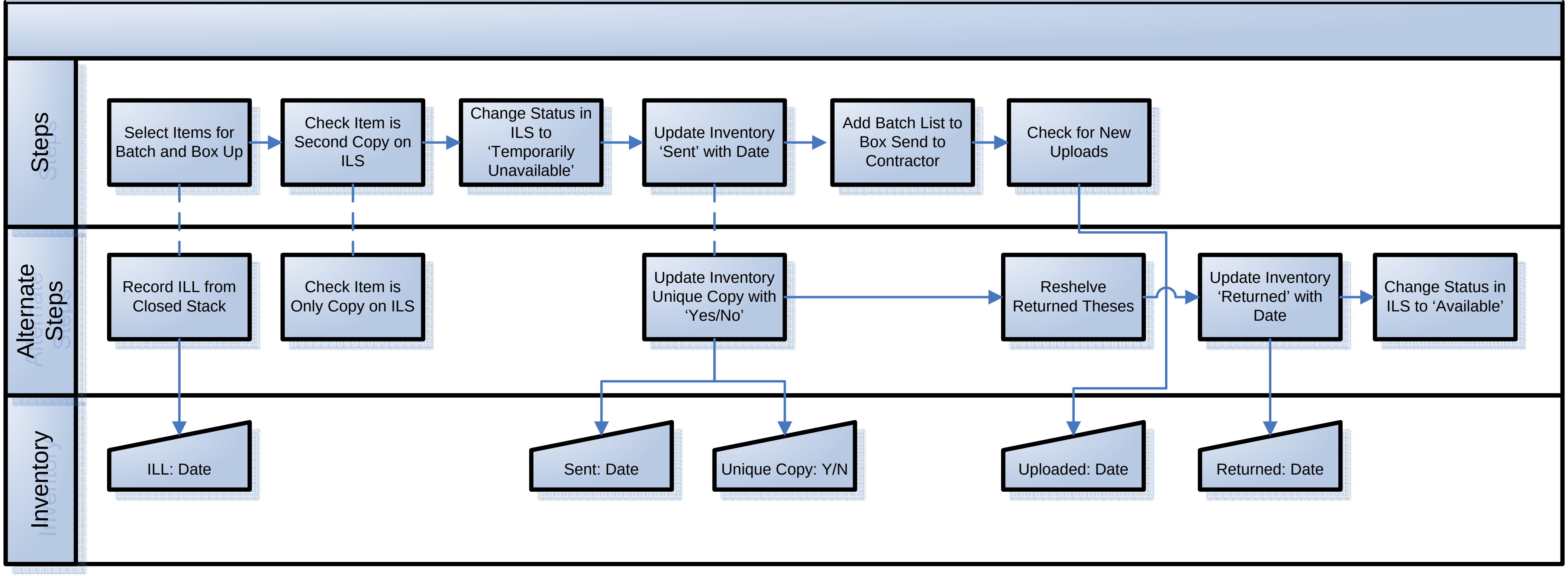


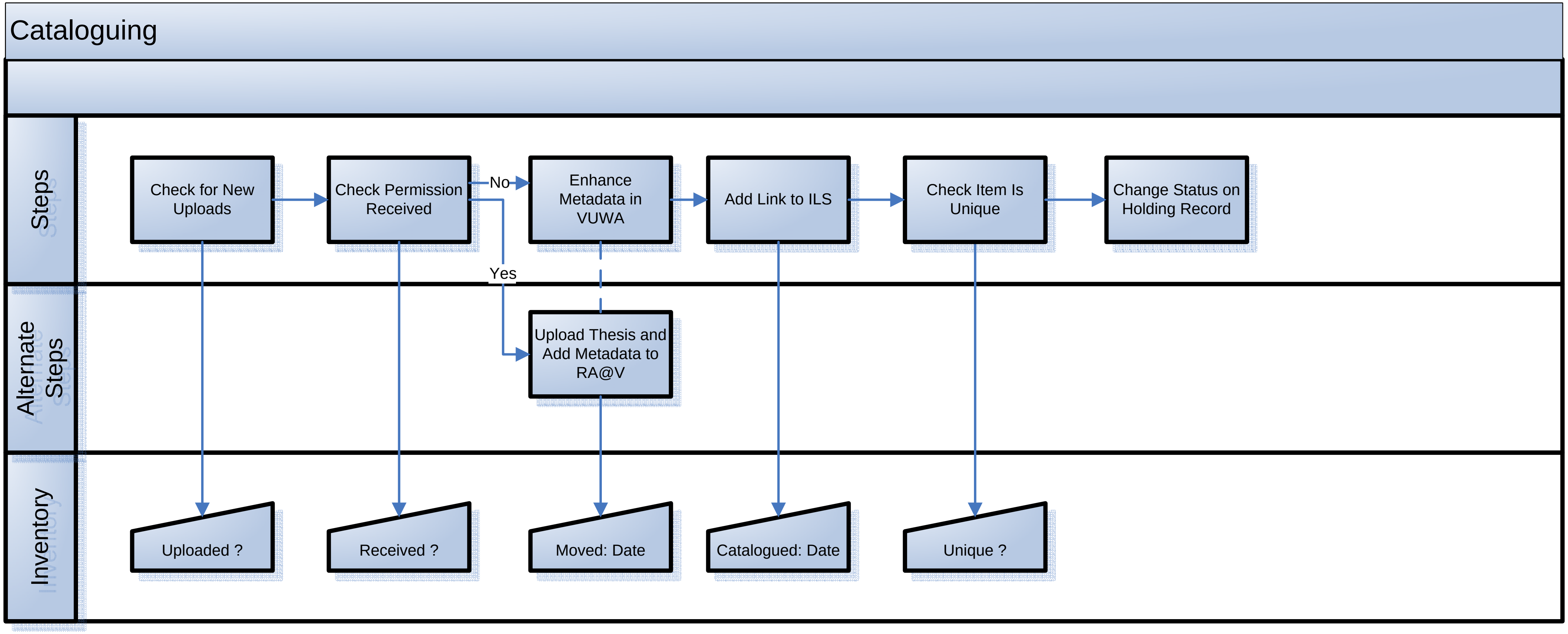


Permissions
Process
Steps
People
Workflow
Permissions Sent
Permissions Received
Permissions Database & VUWA Checked
Implement Permissions Database
Generate & Email Letters
Record Permission in Database
Check New Permissions and Batches Uploaded
IR Coord
Assistant
Assistant
Catalogue Librarian
2A/2B Permission Sent/Received
2A/2B Permission Sent/Received


Description
Process
Update Repositories
Update Catalogue
Steps
Move Theses with Permission to RA@V
Add or Create Metadata in VUWA or RA@V
Remove VUWA Item and Record if Moved to RA@V
Insert Handle in MARC 856 in ILS
People
Catalogue Librarian
Catalogue Librarian
IR Coord
Catalogue Librarian
Workflow
3A/3B Thesis Uploaded/Moved
3A/3B Thesis Uploaded/Moved
4 Catalogued

Thesis Batching


Steps
Alternate Steps
Inventory
Select Items for Batch and Box Up
Check Item is Second Copy on ILS
Change Status in ILS to 'Temporarily Unavailable'
Update Inventory 'Sent' with Date
Add Batch List to Box Send to Contractor
Check for New Uploads
Record ILL from Closed Stack
Check Item is Only Copy on ILS
Update Inventory Unique Copy with 'Yes/No'
Reshelve Returned Theses
Update Inventory 'Returned' with Date
Change Status in ILS to 'Available'
ILL: Date
Sent: Date
Unique Copy: Y/N
Uploaded: Date
Returned: Date


Cataloguing
Steps
Alternate Steps
Inventory
Check for New Uploads
Check Permission Received
No
Yes
Enhance Metadata in VUWA
Add Link to ILS
Check Item Is Unique
Change Status on Holding Record
Upload Thesis and Add Metadata to RA@V
Uploaded ?
Received ?
Moved: Date
Catalogued: Date
Unique ?

Workflow and Inventory Control – Thesis Digitisation

An Access database is being developed to enable the Library to:

1. Generate permission letters
2. Maintain a record of permissions sent and received
3. Record the transfer of theses to (and where appropriate return from) the contractor
4. Maintain a record of items uploaded to the IR
5. Maintain a record of items described in the IR and library catalogue

Data Available

Banner – Data Table

Label	Value
Student ID	196789876
Date of Birth	14/01/1930
Last Name	MASON
First Name	Donald
Email	donald.mason@ispprovider.com
Street1	9 Frances Street
Street2	Waikanae
Street3	
City	Kapiti
Year	1960
Degree	PHD
Degree Desc	Doctor of Philosophy
Major	POLS
Major Desc	Politics
Department	HIPPI
Status	DA
Status Desc	Degree Conferred
Graduation Date	dd/mm/yyyy
Student Level	PG
Student Thesis (title)	Thesis submitted: Building Octagonal Houses

Alumni – Data Table

Label	Value
Student ID	Pre 1990 no match
Date of Birth	
First Name	
Last Name	
Street1	
Street2	
City	
Country	
No Valid Address	
Degree Description	
Email	

ILS – Data Table

Label	Value
Author	
Student Thesis (title)	
Call Number	
Year	
E-version	
Copies	

Requirements

Permission Letter – Data Utilised

Label	Value
First Name	Donald
Last Name	MASON
Street1	9 Frances Street
Street2	Waikanae
Street3	
City	Kapiti
Country	New Zealand
Author	Mason, Donald Ramsay
Student Thesis (title)	Thesis submitted: Building Octagonal Houses
Degree Description	Doctor of Philosophy
Major Description	Politics
Email	donald.mason@ispprovider.com

Research Administrator – Input Data Form

Label	Type	Permission
First Name		Update
Last Name		Update
Author		Update
Student Thesis		Update
Degree Description		Update
Major Description		Update
Department		Update
Uploaded	Date	Update
Permission Sent	Date	Update
Permission Received	Date	Update

Catalogue Librarian – Input Data Form

Label	Type	Permission
First Name		Read
Last Name		Read
Author		Read
Student Thesis		Read
Degree Description		Read
Major Description		Read
Department		Read
Moved	Date	Update
Uploaded	Date	Update
Permission Sent	Date	Update
Permission Received	Date	Update
Catalogued	Date	Update

Lending Librarian – Input Data Form

Label	Type	Permission
First Name		Read
Last Name		Read
Author		Read
Student Thesis		Read
Degree Description		Read
Major Description		Read
Call Number		Read
ILL	Date	Update
Unique Copy	Y/N	Update
Batch Number	Numeric	Update
Thesis Sent	Date	Update
Thesis Returned	Date	Update

Inventory – Data Table

Label	Data	Workflow	Data Match
Student ID	Alumni		
Date of Birth	Alumni		
First Name	Alumni		Yes
Last Name	Alumni		Yes
Author	ILS		Yes
Student Thesis	ILS		
Degree Description	Alumni		
Major Description	Alumni		
Department	Alumni		
Street1	Alumni		
Street2	Alumni		
Street3	Alumni		
City	Alumni		
Country	Alumni		
Email	Alumni		
Call Number	ILS		
E-version	ILS		
Copies	ILS		
ILL	Input		
Supplementary Material	Input		
Unique Copy	Input		
Batch Number	Input	1A	
Thesis Sent	Input	1B	
Permission Sent	Input	2A	
Permission Received	Input	2B	
Uploaded	Input	3A	
Moved	Input	3B	
Catalogued	Input	4	
Thesis Returned	Input	5	