



Australian  
National  
University

# Workflow

Heritage Corpus Construction for Scholarly Research

Ingrid Mason, Deployment Strategist, AARNet  
Roxanne Missingham, University Librarian, ANU




# New curatorial practices...

in corpus construction that support distant reading and data mining through API and programming interfaces can be run in parallel with current curatorial practices that support close reading and online browsing through browser and GUI interfaces.






# What

- Digitising 199 volumes of the Sydney Stock Exchange records
  - Each volume ~300 pages and 100MB sized files
  - Deposit direct into CloudStor direct (by digitisation vendor)
  - Testing data transfer tools
  - Parallel process for deposit into DSpace
  - Mix of printed and handwritten text
  - Breaking up text for transcription and machine learning
- 

# What the..?

- Digital curation knowledge can be reapplied in making large digitised heritage collections available “as data” or as a corpus and existing curation information remains relevant and has contextual value
  - Take opportunities to look at new tools and processes being delivered in the cloud, as there are definitely better and faster ways for digitisation to happen
  - Researchers and custodians are going to become used to interacting with different tools and interfaces in their respective workflows (both involve data handling and processing)
- 

# Project Team

Roxanne Missingham

ANU University Librarian

Erin Gallant & Stephanie Luke

ANU Library Digisquad

Sarah Lethbridge and Maggie Shapley

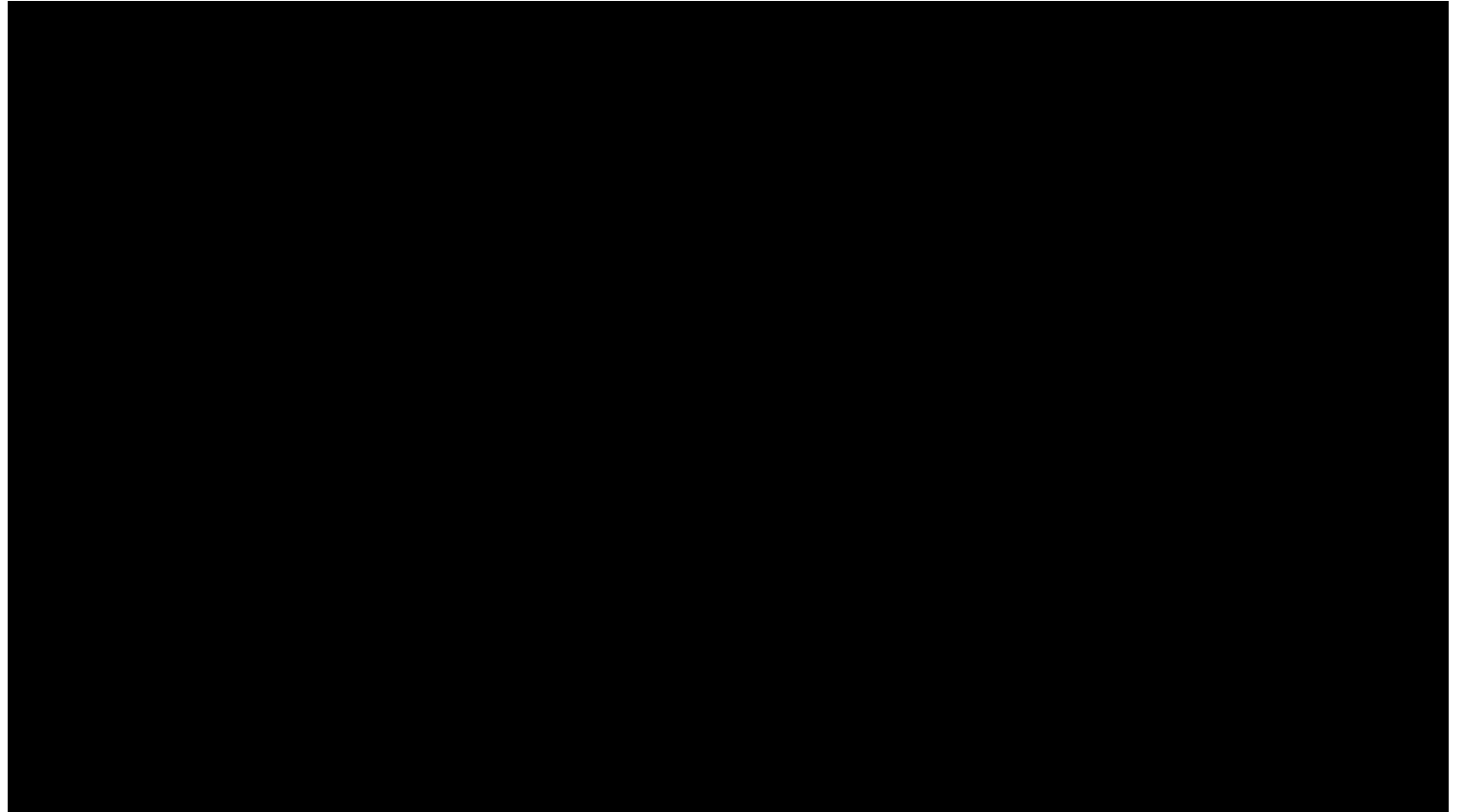
Noel Butlin Archive Archivists

Dr Tim Sherratt

Historian and Hacker

Ingrid Mason

AARNet Strategist





# Digitisation Workflow

Aim: to enable the transfer of files from vendor to library to load directly into cloud storage and into library repository system to be faster and more efficient (by using new approaches and tools).



# Transfer files to cloud storage (CloudStor)

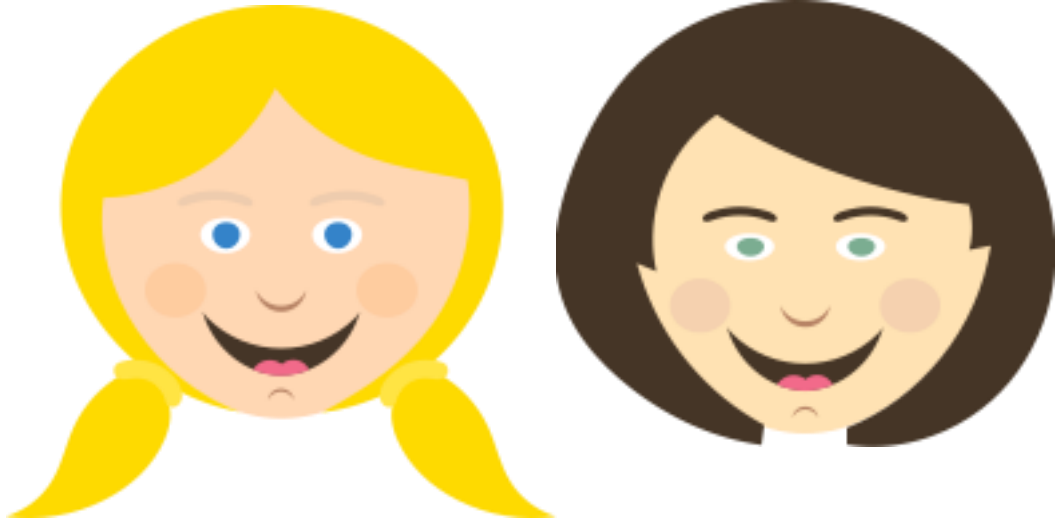
## Transfer from vendor to library

- Used vouchers for the vendor to load files directly into cloud storage 🤖
- Issues with vendor network connection speeds 😬
- Reverted to having external drives being couriered 📦

## Transfer from external drive to library FNS and cloud storage

- Tested library computer and network connection speeds 😊
- Used sync client (off desktop) 😬
- Used WebDAV (off desktop and external drive) 😊
- Used Rocket (off external drive) 😊

DigiSquad



## *Digitisation workflow*

- *Vendor loaded 38 folders. In the end sent a 10TB backup drive with all the remaining folders of files that hadn't been loaded yet (161 folders).*
- *We abandoned the sync client and went to using WebDAV to directly communicate with CloudStor by creating folders and then copying each folder across individually directly from the hard drive. Using WebDAV removed the need to store the files locally (like the sync client does) and we managed to avoid the issues we were having with local disc space on the PCs.*

# Lessons learned: data movement

- End to end digitisation operating online = doable
- Syncing works well with a small number of big files, not well with tons of small files
- WebDAV and Rocket are tools better suited to this type of file transfer process

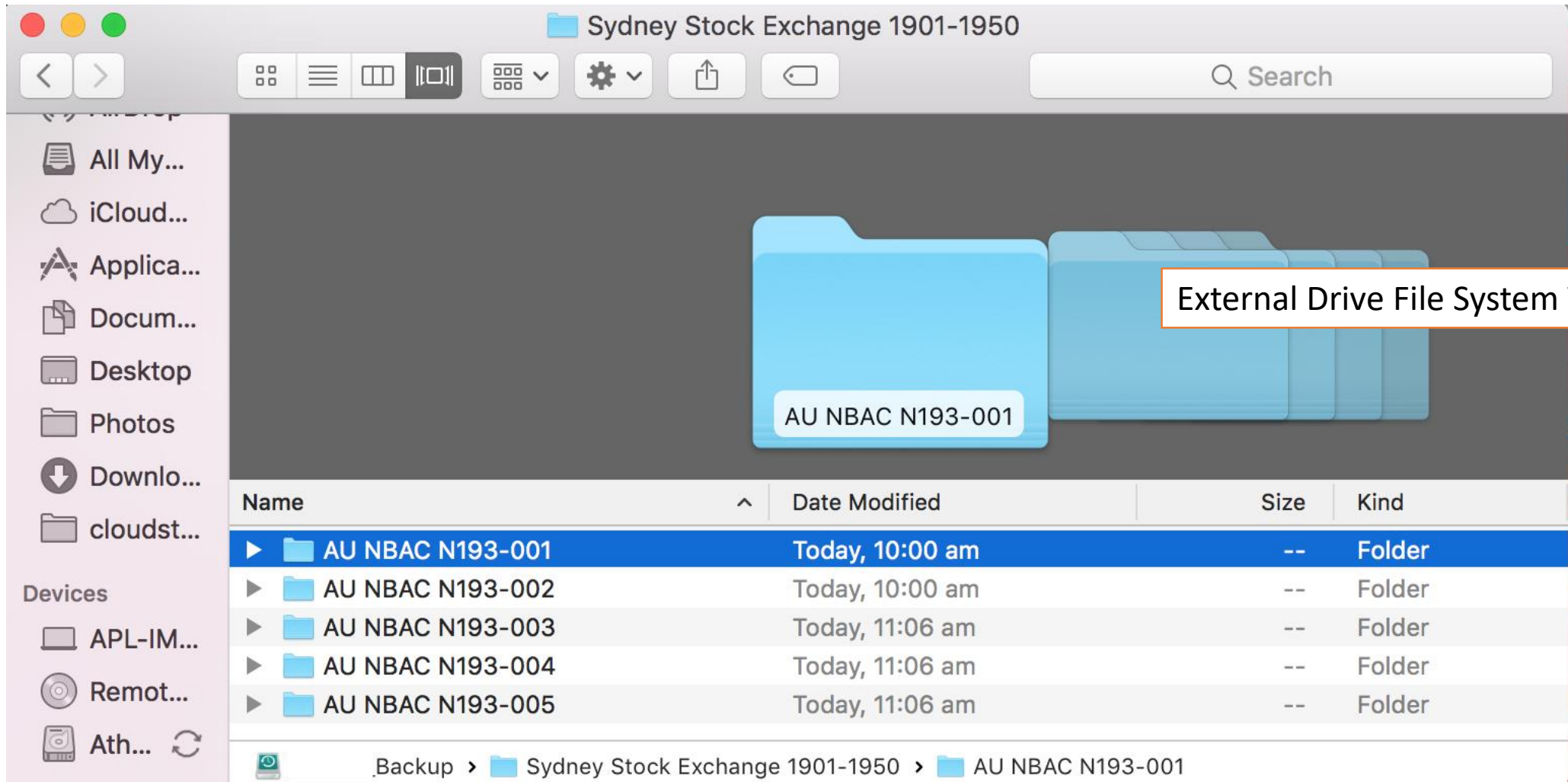
# Tools + Interfaces

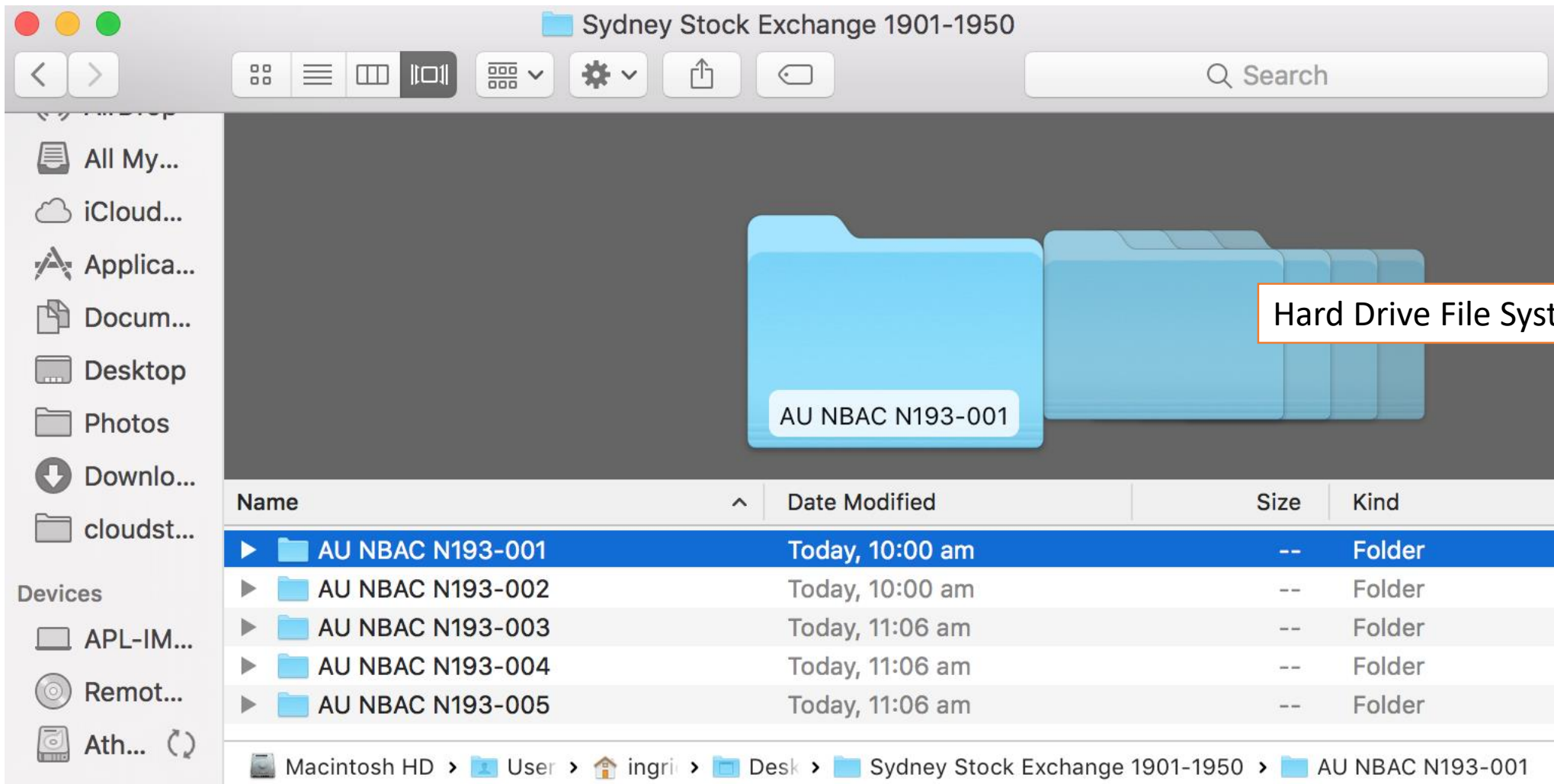
Aim: to enable archivists and librarians to explore a new curatorial processes and arrangement practices and researchers to use computational techniques to query the collection directly and process the data.



# Viewing directories and files

Through old and new system interfaces






The screenshot shows the Cloudstor web interface. On the left is a dark sidebar with navigation options: All files (highlighted with an orange arrow), Favourites, Shared with you, Shared with others, Shared by link, Deleted files, and Settings. The main area shows a breadcrumb path: Shared > ANU-Library > Sydney Stock Exchange 1901-1950 > +. Below the path is a table of files with columns for Name, Size, and Modified. The files listed are folders named 'A Collections' and 'AU NBAC N193-001' through 'AU NBAC N193-008', each with a share icon, a size, and a modification date.


Name	Size	Modified
A Collections	9.8 GB	4 months ago
AU NBAC N193-001	29.1 GB	7 months ago
AU NBAC N193-002	28.2 GB	6 months ago
AU NBAC N193-003	30.9 GB	6 months ago
AU NBAC N193-004	28.2 GB	6 months ago
AU NBAC N193-005	27 GB	6 months ago
AU NBAC N193-006	29.7 GB	6 months ago
AU NBAC N193-007	30.6 GB	6 months ago
AU NBAC N193-008	29.4 GB	7 months ago

Cloud File System View


## Desktop Clients



Windows  
7, 8 and 10




Mac  
10.7+, 64 bit




Linux  
Multiple Distributions



## Mobile Apps



Android




Apple iOS

## Scientific Instrumentation Upload Only Client




Rocke  
Windows  
7, 8 and 10


## Desktop Clients



Windows  
7, 8 and 10




Mac  
10.7+, 64 bit




Linux  
Multiple Distributions

## Mobile Apps



Android



Apple iOS

## Scientific Instrumentation Upload Only Client



Rocke  
Windows  
7, 8 and 10



# Testing out tools and methods

How can digitisation staff move data more efficiently?

# WebDAV

- World Wide **Web** Distributed **A**uthoring and **V**ersioning
- A protocol and an extension of HTTP
- Supports remote access, authoring and movement of files on a server
- File copy and move functions operate on URI (e.g. `http://...`)

# Rocket



**Options** — □ ✕

## Upload Settings

Upload Chunk Test Sizes

<input checked="" type="checkbox"/> 100 KB	<input checked="" type="checkbox"/> 500 KB	<input checked="" type="checkbox"/> 1 MB	<input checked="" type="checkbox"/> 10 MB
<input checked="" type="checkbox"/> 20 MB	<input checked="" type="checkbox"/> 50 MB	<input checked="" type="checkbox"/> 80 MB	<input checked="" type="checkbox"/> 100 MB
<input type="checkbox"/> 200 MB	<input type="checkbox"/> 500 MB	<input type="checkbox"/> 1 GB	<input type="checkbox"/> 1.5 GB

**Run Tests**

Upload Chunk Size 10 MB

Parallel Uploads 4

Data Buffer 4

Max Files Per Chunk 4

Failed Upload Retries 4

**Save Settings**

# Collection arrangement + access

How can researchers processing data access it as a corpus?

Holdings

Quick search

▼ Deposit N193 - Sydney Stock Exchange ...

Item 1 - Sydney Stock Exchange Stock a...

Item 2 - Sydney Stock Exchange Stock a...

Item 3 - Sydney Stock Exchange Stock a...

Item 4 - Sydney Stock Exchange Stock a...

Item 5 - Sydney Stock Exchange Stock a...

194 more...

## Deposit N193 - Sydney Stock Exchange stock and share lists

### Identity area

Reference code	N193
Title	Sydney Stock Exchange stock and share lists
Date(s)	<ul style="list-style-type: none"><li>1901 - 1950 (Creation)</li></ul>
Level of description	Deposit
Extent and medium	199 items

### Context area

Name of creator	<a href="#">Sydney Stock Exchange</a> (1871 - 1987) Administrative history: The Sydney Stock Exchange was formed to allow brokers and traders to trade stocks and bonds for companies listed in New South Wales. It formed an association with the stock exchanges in Adelaide, Melbourne, Brisbane Perth and Hobart called the Australian ... <a href="#">»</a>
-----------------	---

### Content and structure area

Scope and content	These are large format bound volumes of the official lists that were posted up for the public to see - 3 times a day - forenoon, noon and afternoon - at the close of the trading session in the call room at the Sydney Stock Exchange. The closing prices of ... <a href="#">»</a>
-------------------	--

### Conditions of access and use area

Conditions governing access	Researchers must sign an access agreement.
-----------------------------	--

### Access points

Name access points	<ul style="list-style-type: none"><li><a href="#">Sydney Stock Exchange</a> (Creator)</li></ul>
--------------------	---

### Description control area

Dates of creation revision deletion	Entered from deposit description on 25 February 2013.
-------------------------------------	---

### Clipboard

 Add


### Explore

 Reports

 Browse as list

 Browse digital objects

### Export

 Dublin Core 1.1 XML


 EAD 2002 XML

### Related people and organizations

[Sydney Stock Exchange](#)  
(Creator)

Search/Browse metadata


## Sydney Stock Exchange Stock Official List of Prices Current (199) - March 1950 - July 1950

 [Download \(587.45 MB\)](#)

- [Statistics](#)
- [Export Reference to BibTeX](#)
- [Export Reference to EndNote XML](#)

Collections	<a href="#">Sydney Stock Exchange</a>
Title:	Sydney Stock Exchange Stock Official List of Prices Current (199) - March 1950 - July 1950
Author(s):	<a href="#">Sydney Stock Exchange (1871 - 1987)</a>
Date published:	1950
Description:	Large format bound volume of the official lists that were posted up for the public to see - 3 times a day - forenoon, noon and afternoon - at the close of the trading session in the call room at the Sydney Stock Exchange. The closing prices of stocks and shares were entered in by hand on pre-printed sheets.
URI:	<a href="http://hdl.handle.net/1885/147041">http://hdl.handle.net/1885/147041</a>

### Download

File	Description	Size	Format	Image
<a href="#">AU NBAC N193-199.pdf</a>		587.45 MB	Adobe PDF	

Search/Browse metadata and digital object



About

Contribute

Publishing

Policy

Copyright

Contact

Statistics

My Open Research

Home » Archive and Library Collections » Noel Butlin Archives Centre » Sydney Stock Exchange

# Sydney Stock Exchange : [199]

Search Sydney Stock Exchange



Browse Sydney Stock Exchange

Browse by:


The Sydney Stock Exchange was formed to allow brokers and traders to trade stocks and bonds for companies listed in New South Wales. It formed an association with the stock exchanges in Adelaide, Melbourne, Brisbane Perth and Hobart called the Australian Associated Stock Exchanges but remained an independent body. These six stock exchanges amalgamated on 1 April 1987 to form the Australian Stock Exchange Limited (ASX)

Collection's Items (Sorted by Submit Date in Descending order): 1 to 20 of 199



Sydney Stock Exchange Stock Official List of Prices Current (199) - March 1950 - July 1950

Author(s)	Sydney Stock Exchange (1871 - 1987)
Type	Text
Date Published	1950
Date Created	March 1950 - July 1950

RECENT SUBMISSIONS 

Sydney Stock Exchange Stock Official List of Prices Current (199) - March 1950 - July 1950

Sydney Stock Exchange Stock and Share Lists (198) - 03 April 1950 - 30 June 1950

Sydney Stock Exchange Stock and Share Lists (197) - 04 January 1950 - 31 March 1950

Sydney Stock Exchange Stock and Share Lists (196) - 04 October 1949 - 22 December 1949

Sydney Stock Exchange Stock and Share Lists (195) - 01 July 1949 - 30 September 1949

Search/Browse metadata and digital object

Statistics

## Getting file lists

```
In [38]: # Ok let's initiate the client.
client = wc.Client(options)
```

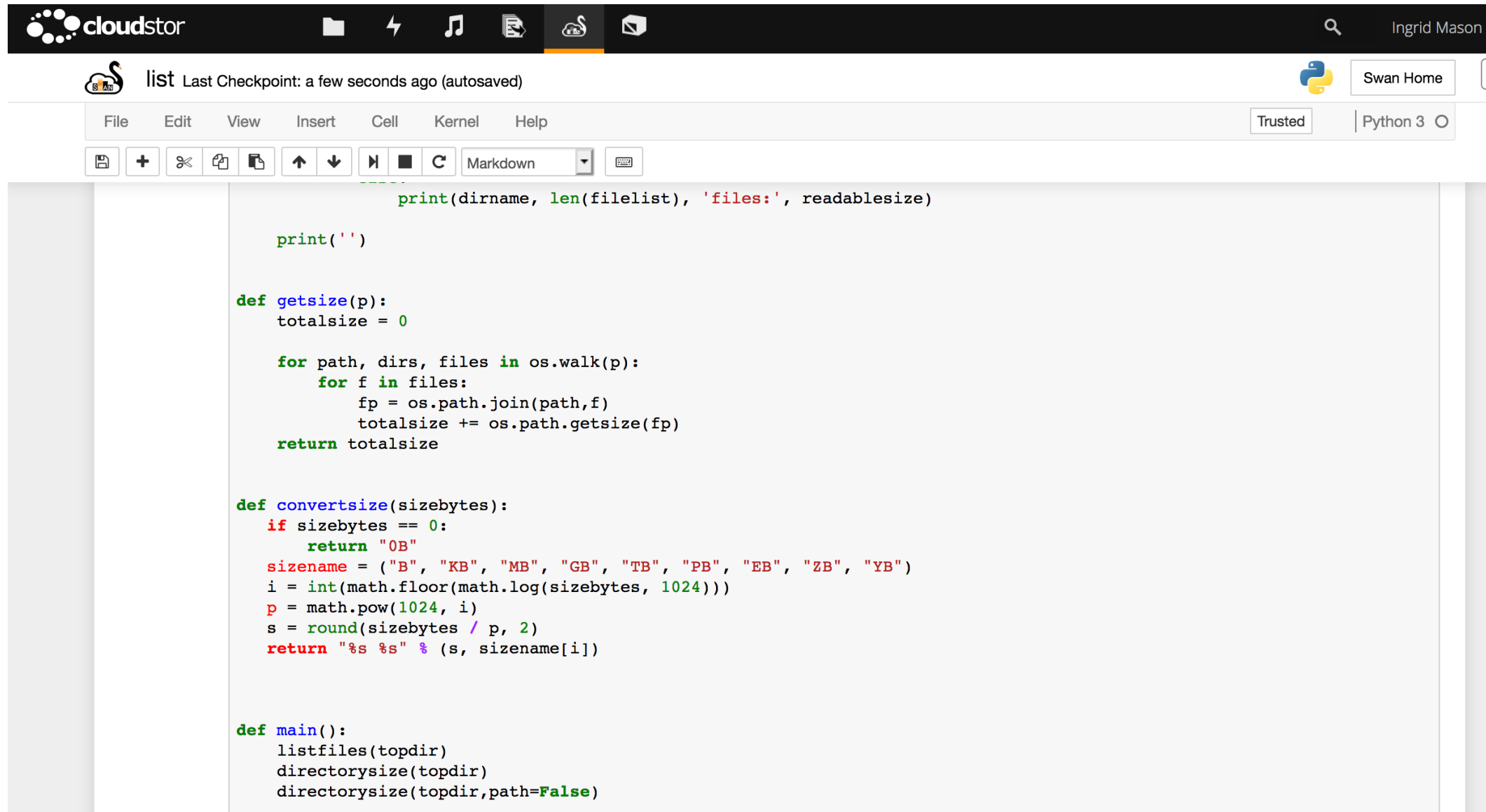
```
In [39]: # Use .list() to get a list of resources in the directory
# In this case it's a list of subdirectories
dirs = client.list('Shared/ANU-Library/Sydney Stock Exchange 1901-1950/')
# For some reason the parent directory is included in the list, let's filter it out
dirs = [d for d in dirs if d[:2] == 'AU']
```

```
In [40]: # Loop through all the subdirectories and use .list() again to get all the filenames
details = []
summary = []
for d in dirs:
    files = [f for f in client.list('Shared/ANU-Library/Sydney Stock Exchange 1901-1950/{}'.format
(d)) if f[:1] == 'N']
    print('{}: {} files'.format(d, len(files)))
    # Save the details for each subdirectory
    summary.append({'directory': d, 'number': len(files)})
    for f in files:
        path = 'Shared/ANU-Library/Sydney Stock Exchange 1901-1950/{}'.format(d, f)
        # This slows things down a lot, so disable for now
        # info = client.info(path)
        info = {}
        info['name'] = f
        info['directory'] = d
        info['path'] = path
        # print(info)
        details.append(info)
    time.sleep(0.5)
```

```
AU NBAC N193-001/: 303 files
AU NBAC N193-002/: 312 files
AU NBAC N193-003/: 345 files
AU NBAC N193-004/: 312 files
AU NBAC N193-005/: 305 files
AU NBAC N193-006/: 334 files
AU NBAC N193-007/: 349 files
```



CloudStor Notebook here... showing Python sub-directory and file inventorying techniques, based on Dr Tim Sherratt's Jupyter notebooks housed in GitHub (accessing files in CloudStor remotely) by public token or authentication/authorisation.



The screenshot displays a Jupyter Notebook interface. At the top, there is a dark header with the 'cloudstor' logo on the left and a search icon and the name 'Ingrid Mason' on the right. Below this is a white bar with the notebook title 'list' and a status message 'Last Checkpoint: a few seconds ago (autosaved)'. On the right side of this bar, there is a Python logo and a button labeled 'Swan Home'. The main interface features a menu bar with 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', and 'Help'. To the right of the menu bar, there is a 'Trusted' label and a dropdown menu set to 'Python 3'. Below the menu bar is a toolbar with various icons for file operations and execution. The central area of the notebook contains a code cell with the following Python code:

```
print(dirname, len(filelist), 'files:', readablesize)

print('')

def getsize(p):
    totalsize = 0

    for path, dirs, files in os.walk(p):
        for f in files:
            fp = os.path.join(path,f)
            totalsize += os.path.getsize(fp)
    return totalsize

def convertsize(sizebytes):
    if sizebytes == 0:
        return "0B"
    sizename = ("B", "KB", "MB", "GB", "TB", "PB", "EB", "ZB", "YB")
    i = int(math.floor(math.log(sizebytes, 1024)))
    p = math.pow(1024, i)
    s = round(sizebytes / p, 2)
    return "%s %s" % (s, sizename[i])

def main():
    listfiles(topdir)
    directorysize(topdir)
    directorysize(topdir,path=False)
```

# In the world of linguistics...

who are well practised at corpus construction, there are varying views on what a corpus is and how formally arranged and constituted (in terms of content) it is.




# Lessons earned: corpus construction

- Consistent directory and file naming aligned with the archival description (readability).
- Corpus arranged for flexible access i.e. targeted, sequential, sub-setting or batch processing (systematic).
- Contextual archival documentation and dual discovery modes (sense-making).

# Data curation considerations

- Is there an research support guidance info to support “data intensive researchers”?
- How does any readme information for the corpus and dataset provided align with what is available online already?
- How do data users cite and acknowledge the corpus or the data derived from the corpus?
- Should a DOI be put into the metadata records in that, so that citation can be programmatically tracked?

# We're still going...

We're onto our second phase, this involves unlocking the data (printed and handwritten) to enable a derived dataset to be generated. DH & tech enthusiasts watch this space  <https://github.com/wragge/sydney-stock-exchange> or for digitisation and digital curators keep across @ANULibrary and @AARNet news.