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) 😊
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
External Drive File System View
Cloud File System View
Testing out tools and methods

How can digitisation staff move data more efficiently?
WebDAV

• World Wide **Web Distributed Authoring and Versioning**
• 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://...)

Collection arrangement + access

How can researchers processing data access it as a corpus?
Deposit N193 - Sydney Stock Exchange stock and share lists

Identity area
- Reference code: N193
- Title: Sydney Stock Exchange stock and share lists
- Date(s): 1901 - 1950 (Creation)
- Level of description: Deposit
- Extent and medium: 199 Items

Context area
- Name of creator: Sydney Stock Exchange (1871 - 1887)
- 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...

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...

Conditions of access and use area
- Conditions governing access: Researchers must sign an access agreement.

Access points
- Name access points: Sydney Stock Exchange (Creator)

Description control area
- Dates of creation revision deletion: Entered from deposit description on 25 February 2013.
Search/Browse metadata and digital object
Sydney Stock Exchange: [199]

Search Sydney Stock Exchange

Browse Sydney Stock Exchange

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

Search/Browse metadata and digital object
Getting file lists

```python
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
        try:
            info = client.info(path)
            info = {}  # info = client.info(path)
            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.

```python
print(dirname, len(filelist), 'files:', readableSize)

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"
    filesize = ("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 "{} {}".format(s, filesize[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.