University Libraries Enabling Cultural Data to Flow

Ingrid Mason - AARNet, Michael McGuinness - Griffith University
The Twitters
Workshop Agenda

2.30pm-4.30pm  Data Handling/Infrastructure - Ingrid Mason
   - Introduction to session
   - Case studies and group exercises
   - Data movement - tools/services
   - Next steps

3.30pm-4.00pm  Break
Workshop Agenda

4.30pm-5.30pm Case Study - Prosecution Project - Michael McGuinness

- Liaison with cultural institutions for collection access
- Reuse of digitised material and arrangement for digitisation
- Transfer of digitised material (different approaches)
- Transcription platform & 360 degree data sharing (RDS C&C)
Introduction to Session
Library Support Role

Enabling Cultural Data to Flow

- Data Handling
- Infrastructure
Covering Today - Data Handling/Infrastructure

- Librarians support role (discussion)
- Speed test (individual exercise)
- Data handling (definition)
- Data packaging (definition)
- Case studies (group exercise)

- Networks (description)
- File sizes and network speeds (examples)
- Syncing and sending (what/why)
- Encryption (what/why)
What can university library collections and support staff in university libraries do to enable cultural data to be more accessible and to assist in getting cultural data to flow?
Quite a bit..
What Can University Libraries Do?

Learn more about data curation (data packaging and handling)
Share knowledge around digitisation and data management
Offer support for making data accessible and moving data on the NREN
Offer library collections “as data” for research, teaching and learning
Data Handling

Move the data this way
Speed Test!

Is your data moving at high speed? If not/why not?
Speed Test

Use your phone, computer and tablet
Use different applications on each device
Use wifi vs ethernet at home or work (and compare)
Test at different times of the day or night
Test with different ISPs

Use different speed testing tools and calculators
Test with upload and download
Note the similarities or differences
Note the test times
Note the different speed metrics: line, download, upload
Data Handling & Packaging
Data Handling

Merriam Webster definition of handling:

- the act of touching, feeling, holding, or moving something
- the way that someone deals with a person, event, situation, etc.
- the act or process of packing and shipping something to someone (such as a customer)
Data Handling

Data storage and retrieval
Data management
Data access or supply
Data Packaging

Merriam Webster definition of packing:

the action or process of packing something; also : a method of packing

material (as a covering or stuffing) used to protect packed goods (as for shipping)
Data Packaging

Data curation

Data reuse

**FAIR** (findable accessible interoperable reusable)

Metadata
Data Curation

Merriam Webster definition of curator:

*One who has the care and superintendence of something;*
## Data Handlers

- Data librarians
- Repository managers
- Data curators
- Data analysts
- Data scientists
- Researchers
- Lecturers
- Students
Data Handling

Capacity to:

- Select, package, and transfer data
- Use a data storage and transfer tool (e.g. CloudStor)
- Complete a technical assessment

Skills, tools and services to:

- Arrange data - what data is selected and how is the data packaged up
- Supply data - how is the data being made available (in situ or via transfer)
- Share data - who needs to access to the data (and why?)
Data Handling Workflow

- Capture/store data
- Load & transfer data via cloud storage
- Ingest data into Virtual Lab, Electronic Notebook, Data Repository etc
- Load & transfer data via cloud storage
- Capture/store data
Format Conversion - Institution A > University

**Workflow**

- PM identified
- Format conversion
- DM transferred how?
- Data transfer via:
  - A.media
  - B.network
  - C.service
- DM received

**Steps**

- PM identified
- Format conversion
- DM transferred how?
- Data transfer via:
  - A.media
  - B.network
  - C.service
- DM received

**Parties**

- Researcher
- Institution
- Researcher
- Institution
- Researcher
- Institution

Material is already digitised. Data transferred via media (hard-drive) from Institution A to University.
Format Conversion - InstitutionB > University

Material is already digitised. Data transferred via API (FTP) from Institution B to University.
Case Studies & Group Exercises
Case Study Exercise

What role can university libraries play in enabling cultural data to flow?

What next steps can be taken by university libraries to assist in getting cultural data to flow?

How “chunky” or “big” cultural data may be transferred and access to it enabled by the university library?

What tools and services are available to university library staff to support cultural data handling?
Scenario 1

A CRIMINAL Life
Clue: There’s no escape without a ticket.
Scenario 2

CITY SPIRIT
Clue: Mind your language!
What YOU Do?
Case Studies (based on real examples)

Scenario 1: A Criminal Life

Scenario 2: City Spirit
Data Handling Workflow

- Capture/store data
- Load & transfer data via cloud storage
- Ingest data into Virtual Lab, Electronic Notebook, Data Repository etc
- Load & transfer data via cloud storage
- Capture/store data
Data Handling Workflow

Digital assets already in digital repository hosted locally or remotely.

Local data handling (within an organisation)

Shared data handling (within or across organisations)
Shared Data Handling

Cloud data storage and movement tools
CloudStor, Local, OneDrive, Google Drive, DropBox
End-to-End Process

Capture/store data

Load & transfer data via CloudStor

Ingest data into Omeka, Gephi, Neo4j, ArcGIS etc

Load & transfer data via CloudStor.

Capture/store data

Local DSpace

Cloud Storage

Sender/Share

Cloud Storage

Local App

Cloud Storage

Sender/Share

Local DSpace
Case Studies (based on real examples)

Scenario 1: A Criminal Life

Scenario 2: City Spirit
The Big Reveal...
Case Studies (the real ones)

Scenario 1: Prosecution Project

Scenario 2: QuakeBox Project
Data Movement
Elephants

Come in different shapes and sizes.
Move slowly and quickly.
Move together as a group and apart by themselves.
So does data.
The network has been purpose built. "Elephant" Flows
Networks
Computer Networks

Networks can be dedicated or shared.

Networks can be designed with different topologies e.g. spurs, rings, stars.

Networks can carry different levels and types of traffic e.g. Mbps, Gbps.

Networks are affected by interference e.g. microwaves, lift motors, air-conditioning fans.

Networks are physically comprised of cables or wireless media.

Organisations restrict traffic into or out-of their networks via firewall.
Wifi

The conference wifi is provided by Multimedia.

The conference wifi operates at Mbps whereas NRENs can operate at Gbps.

Wifi has no “express lane”.

Wifi privileges no user over another.
Network Capacity

Firewalls can be like manual “toll booths” and that can slow traffic down.

Firewalls can be like multilane freeways with automated toll gathering.

Science Demilitarised Zone (Science DMZ) is a dedicated network connection that bypasses the firewall “freeway” (configured for data transfer).

Building structures and the physical network have an impact on capacity.

Test nodes are provided to IT staff to undertake scheduled tests to aid in detecting network “bottlenecks”.
National Research & Education Networks

NRENs in Australian and New Zealand: AARNet & REANNZ

International community e.g.

Internet2 - USA (national) and CalREN (Calif state)

Myren - Malaysia

GÉANT- pan European

GARR - Italy
NRENs

Mission “to provide high speed resilient connection services” especially to meet big data movement. (Building Tomorrow’s Big Science Networks, TNC 2017)

Currently building to support:

- Exponential demand for bandwidth
- Expanding requirements (beyond big data movement) e.g. complex workflows
- More nuanced control and management of network
NRENs
Australia’s National Broadband Network

<table>
<thead>
<tr>
<th>Download (Mbps)</th>
<th>Upload (Mbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
</tr>
<tr>
<td>50</td>
<td>20</td>
</tr>
<tr>
<td><strong>100</strong></td>
<td><strong>40</strong></td>
</tr>
</tbody>
</table>
Australia’s NREN

How long does it take to transfer 1 Petabyte of data across various speed networks?

- 100Mbps: 3.4 Years
- 1Gbps: 4 Months
- 10Gbps: 10 Days
- 100Gbps: 1 Day

(Note: Actual speeds may vary for a range of infrastructure reasons)
New Zealand’s NREN
"The difference between a week and a day

The work is understandably data-intensive, with a specific requirement to be able to give access to this data to local and international colleagues. It wouldn’t be possible on a traditional ISP network. This level of data transfer requires an ultra high speed fibre connection. Up until recently, it would still take weeks to move the data observed by the telescopes. However, since the 10Gb/s link from Warkworth was installed, Stuart has been able to transfer these same experiment data sets in under a day. “The difference between a week and a day is huge. It makes a big difference to my work and what I’m capable of achieving,” he says.”
File Sizes and Network Speeds
Technical Assessment

Consider the scale of the data when planning data handling and other factors:

- Number of files: 10, 100, 1000, 10,000
- File sizes: KB, MB, GB
- Network speeds: Kbps, Mbps, GBps
- Networks: commodity/domestic or NREN
- Service point: catalogue, API, download, FTP (take away)
- Remote access: researcher accesses data remotely (big or sensitive)
Speed Test - Downloading from Home

100 megabytes (100MB) = ~7 m

Download speed: ~3.38 Mbps

100 gigabytes (100GB) = ~119 h 21 m

(Upload speed: 0.50 Mbps)

100 terabytes (100TB) = ~119357 h 38 m
Speed Test - Downloading at a Conference

100 megabytes (100MB) = ~17 s

Download speed: ~54 Mbps

100 gigabytes (100GB) = ~4 h 47 m

100 terabytes (100TB) = ~4793 h 1 m
Speed Test - Downloading on the NBN

100 megabytes (100MB) = ~8 s

100 gigabytes (100GB) = ~2 h 26 m

100 terabytes (100TB) = ~2444 h 26 m

Download speed: ~100 Mbps
Speed Test - Downloading from Work

100 megabytes (100MB) = \(~1\) s

100 gigabytes (100GB) = \(~24\) m

100 terabytes (100TB) = \(~393\) h

Download speed: \(~680.52\) Mbps

(Upload speed: 844.72 Mbps)

Laptop and network connection have not been optimised for data transfer.
Tools and Services

ScienceDMZ

CloudStor

Collections plugin
Science DMZ

A portion of the network, *built at or near the campus or laboratory’s* local network perimeter.

*Designed* such that the equipment, configuration, and security policies are optimised *for high-performance scientific applications and data movement.*

*A separate route* for data than that for general-purpose business systems or “enterprise” computing.
Data Access and Transfer Tools

Make data accessible (conditionally) e.g. data access is controlled (open, by password)

Make data transferrable (conditionally) e.g. data transferred (from a A to B) is defined as (one off, recurrent, syncing)

Make data secure e.g. data is encrypted (end to end)
CloudStor
Synching
Sharing

- Shared with you
- Shared with others
- Shared by link
CloudStor Collections Plugin
CloudStor Collections Plugin

Enables researchers to easily collate, annotate, package and share groups of files and metadata, and is designed to support research data packaging for the long tail of research groups (MB-TB-GB scale)

Features of v1.1

- Select any destination for data package transfer.
- Readme file (RDFa) machine+human readable options.
- Select different metadata model for different ingest targets.
Encryption (at rest & in transit)
Encryption (at rest & in transit)
Next Steps
You Can

Take a look at Always Already Computational - Collections as Data IMLS project and have a go at using the Collections plugin once it is launched in CloudStor (data literacy).

Create your own "how to" video on how to effectively use the network and storage services to move data for your university community (infrastructure literacy).

Take a look at the RDS Cultures & Community project as an exemplar collaboration between humanities researchers in a university collaborating with government archives (cultural data flow).
THANK YOU
Case Study - Prosecution Project along with the Cultures and Community Project (Open API)

Participating Institutions:

- VicNode
- Griffith University
- NLA
- Queensland State Archive
- Deakin University
- INTERSECT
- Aarnet

Project funded by: RDS Research Data Services
Workshop Agenda

4.30pm-5.30pm  Case Study - Prosecution Project - Michael McGuinness
  - Liaison with cultural institutions for collection access
  - Reuse of digitised material and arrangement for digitisation
  - Transfer of digitised material (different approaches)
  - Transcription platform & 360 degree data sharing (RDS C&C)
About Me

Michael McGuinness
Business Analyst
eResearch Services
Griffith University
@M_J_Mcguinness
Liaison with cultural institutions for collection access
- The barriers
- Negotiations required
- Old Bailey
- Legacy Systems
Archives & project partner locations

- State Archives
- eRSA
- Tasmanian Archive and Heritage Office
- The Prosecution Project
- Griffith University
- National of Library of Australia
Reuse of digitised material and arrangement for digitisation
Different Priorities
Record storage types
Records Lost
Pressures
Transfer of digitised material (different approaches)
- Email
- FTP
- Hard Drives
- AARNET
  - Cloudstor
Workflow of adding images to PP

1. From Archive → Local Archive
2. Send/Share
3. Local File Storage → Local File Storage
4. Move on to Network
5. Network Storage
6. Load
7. Application Overnight process
8. Assign / Queue Images
9. Transcribe → QA
10. QA
11. Database
12. Data Storage and Movement
13. Application

Griffith University
Prosecution Project

#THETA2017
Transcription platform & 360 degree data sharing (RDS C&C)
- Metadata
- Archives Metadata
- Advisory Board
- Technical Advisory Board
# Metadata Fields to Share

<table>
<thead>
<tr>
<th>PP</th>
<th>TAHO</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Def_fname</td>
<td>• Title</td>
</tr>
<tr>
<td>• Def_surname</td>
<td>• Description</td>
</tr>
<tr>
<td>• offence status</td>
<td>• Publisher</td>
</tr>
<tr>
<td>• trial_date</td>
<td>• Date</td>
</tr>
<tr>
<td>• trial_id</td>
<td>• Type</td>
</tr>
<tr>
<td>• trial_place</td>
<td>• Format</td>
</tr>
<tr>
<td>• Verdict</td>
<td>• Identifier</td>
</tr>
<tr>
<td></td>
<td>• Source</td>
</tr>
<tr>
<td></td>
<td>• Language</td>
</tr>
</tbody>
</table>
**DC elements (TAHO -> PP)**

<table>
<thead>
<tr>
<th>Metadata Element</th>
<th>Data Type</th>
<th>Data Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc.title</td>
<td>String</td>
<td>Register of Indictments</td>
</tr>
<tr>
<td>dc.publisher.secondary</td>
<td>String</td>
<td>Prosecution Project</td>
</tr>
<tr>
<td>dc.publisher.primary</td>
<td>String</td>
<td>LINC Tasmania</td>
</tr>
<tr>
<td>dc.publisher.uri</td>
<td>URI</td>
<td><a href="http://prosecutionproject.griffith.edu.au/oai/">http://prosecutionproject.griffith.edu.au/oai/</a></td>
</tr>
<tr>
<td>dc.identifier.series</td>
<td>Identifier</td>
<td>AD875</td>
</tr>
<tr>
<td>dc.identifier.item</td>
<td>Identifier</td>
<td>AD875/1/1</td>
</tr>
<tr>
<td>dc.identifier.part</td>
<td>Identifier</td>
<td>AD875/1/1-1</td>
</tr>
<tr>
<td>dc.identifier.subpart</td>
<td>identifier</td>
<td>AD875/1/1-1-1</td>
</tr>
<tr>
<td>URI</td>
<td></td>
<td><a href="https://linctas.ent.sirsiDynix.net.au/client/en_AU/tas/search/detailnonmodal/ent$002f$002fARCHIVES_DIGITISED$002f0$002fARCH_DIGITISED$002f%5C">https://linctas.ent.sirsiDynix.net.au/client/en_AU/tas/search/detailnonmodal/ent$002f$002fARCHIVES_DIGITISED$002f0$002fARCH_DIGITISED$002f\</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Orginal Document</td>
</tr>
</tbody>
</table>
**Custom PP Data DC elements** (PP→ TAHO)

<table>
<thead>
<tr>
<th>Metadata Element</th>
<th>Data Type</th>
<th>Data Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>pp.trial.name</td>
<td>String</td>
<td>Trial of...</td>
</tr>
<tr>
<td>pp.trial.identifier</td>
<td>Identifier</td>
<td>12345</td>
</tr>
<tr>
<td>pp.trial.offence</td>
<td>Controlled term</td>
<td>Murder</td>
</tr>
<tr>
<td>pp.trial.date</td>
<td>Time/date format</td>
<td></td>
</tr>
<tr>
<td>pp.trial.placename</td>
<td>Controlled term</td>
<td></td>
</tr>
<tr>
<td>pp.trial.verdict</td>
<td>Controlled term</td>
<td></td>
</tr>
<tr>
<td>pp.person.firstname</td>
<td>String</td>
<td></td>
</tr>
<tr>
<td>pp.person.lastname</td>
<td>String</td>
<td></td>
</tr>
<tr>
<td>pp.rights.licence</td>
<td>Controlled term</td>
<td></td>
</tr>
</tbody>
</table>
Trends and Drivers

Government Initiatives
Open Science
Unlocking Archives
Citations
Data Sharing
Publications
Citizen Science
Data Access
Crowd Sourcing
Open Data

Environment

GLAM
National Archives of Australia (NAA)
National Library Australia (NLA)
Tasmanian Archives & Heritage Office (TAHO)
Queensland State Archives (QSA)
GLAM PEAK
Atlas of Living Australia (ALA)

Technologies & Standards
OAI-PMH
EAD
EAC-CPF
METS
RDF
Dublin Core
Schema.org
FOAF
JSON
NER & UNI
XML

Research & Other Entities
Family Search
Alveo
Founders & Survivors
The Prosecution Project
HUNI
Ancestry.com
London Lives
Find My Past
Digital Panopticon
e-Ark
Old Bailey Online
# Options: Standards – Transport Layers

<table>
<thead>
<tr>
<th>Method</th>
<th>Pro</th>
<th>Con</th>
</tr>
</thead>
<tbody>
<tr>
<td>OAI-PMH</td>
<td>• TAHO currently use to allow Trove to harvest their data</td>
<td>• Custom setup</td>
</tr>
<tr>
<td></td>
<td>• Is a proven standard used by archives around the world</td>
<td>• Need resources with knowledge of standard</td>
</tr>
<tr>
<td>Resource Sync / Site Map</td>
<td>• Used by Google to produce detailed search results</td>
<td>• TAHO site map is incredibly large and the web-server may crash if you try to read it all in one go.</td>
</tr>
<tr>
<td></td>
<td>• Has a notification mechanism</td>
<td>• Further investigation required as to scalability</td>
</tr>
<tr>
<td></td>
<td>• Low barrier to entry</td>
<td></td>
</tr>
</tbody>
</table>
Proposed Solution/Pilot
Prosecution Project & TAHO

OAI-PMH

Dublin Core
Current State

Prosecution Project (PP)

- TASSC Records
- Images
- Transcription Data

Send MetaData (Manually)

TAHO

- TAHO Court Records
- MetaData
- Images

Send Image Files (Manually) With Image ID's

Michael McGuinness, Griffith University, 27/03/2017
Future Uses

NER workflow solution

Connection to a UNI Solution

Image Sharing

#THETA2017
Poster Reveal

#THETA2017
In Australia, we’re enabling greater access to archives

Oh, we know archives.

Archives across Australia hold data and metadata that is useful to other archival institutions. When archives identify related data, documents and records in an external repository, they typically collaborate with the other archive to further enhance their own collection.

So how do archival institutions share their information? Well, up until now they did it via email, FTP, web-based file sharing services, and in some instances, posting external hard drives. Researchers engage with archives in a similar manner. Most of the interaction is manual and quite often requires a researcher to physically visit the archive.

While these file sharing solutions have achieved the desired outcome between the parties, they can be inefficient, cumbersome and don’t promote broader sharing of data and information. That’s why we’re working on a project to share records across multiple institutional repositories.

How are we doing this?

Alongside stakeholders, such as National Library of Australia, National Australian Archives, Tasmanian Archives and Heritage Office (TAHO) and Queensland State Archives, we’re working on the Open API Project through a program of work managed by eRSA.

The pilot project aims to operationalise a national, sustainable and scalable API standard that will allow data (and metadata) sharing and transfer between the Prosecution Project and TAHO. TAHO is a part of LINC Tasmania. They collect, manage and preserve Tasmania’s cultural and documentary heritage, including State Government records.

The draft design of an Open API Standard will be released in May 2017.
Roadmap to the Open API

Under Trends & Drives, the shaded items have been determined to be affecting the stakeholders and the project is working to assist in resolving these. The figure shows the total environment, the items shaded in light green are the institutions consulted. The major stakeholders are indicated in dark green. These have been selected due to the maturity of their online presence and their already existing strong working relationships.
Prosecution Project: paving the way

The Prosecution Project is a major undertaking, based at Griffith University and funded by the Australian Research Council, which has been investigating the history of the criminal trials in Australia. This project has been funded for five years since 2013. An outcome of the Prosecution Project is an online repository of Australian criminal trial records from 1850-1960. The collection includes data drawn from original court registers, court calendars, trial briefs and police gazettes.

Basically, the goal of the Open API Project is to make data from the Prosecution Project available in TAHO, and vice versa. The API will facilitate the transfer of data through Metadata Endpoints.

The Solution

The outcome sought is to replace the current manual methods with automatic workflows that result in permanent links between objects held at the two repository/databases. See concept pictured left, “360 Datasharing”.

The two workflows below indicate the current state of the workflow and the future state once API is enabled.
Collaboration, it’s key to everything we do.

Visit our Cultures & Community Project  www.ersa.edu.au/cultures-community-project
Visit our Prosecution Project  www.prosecutionproject.griffith.edu.au
Contact us via email at  comms@ersa.edu.au
THANK YOU
Questions?
Griffith UNIVERSITY
Workshop Summary

- **Data Handling/Infrastructure - Ingrid Mason**
  - Introduction to session
  - Case studies and group exercises
  - Data movement - tools/services
  - Next steps

- **Case Study - Prosecution Project - Michael McGuinness**
  - Liaison with cultural institutions for collection access
  - Reuse of digitised material and arrangement for digitisation
  - Transfer of digitised material (different approaches)
  - Transcription platform & 360 degree data sharing (RDS C&C)