#THETA2017

University Libraries Enabling Cultural Data to Flow

Ingrid Mason - AARNet, Michael McGuinness - Griffith University

Ingrid Mason

Deployment Strategist

ingrid.mason@aarnet.edu.au

@1n9r1d

Data Handling #beyondsneakernet @AARNet
Prosecution Project #digitalhumanities @ProsecutionP @Griffith_Uni
C&C Project @eResearchSA
THETA Conference #THETA2017 @thetaconference

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


• Data Handling
• Infrastructure
Library Support Role
Enabling Cultural Data to Flow
© AARNet Pty Ltd |
aarnet

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..

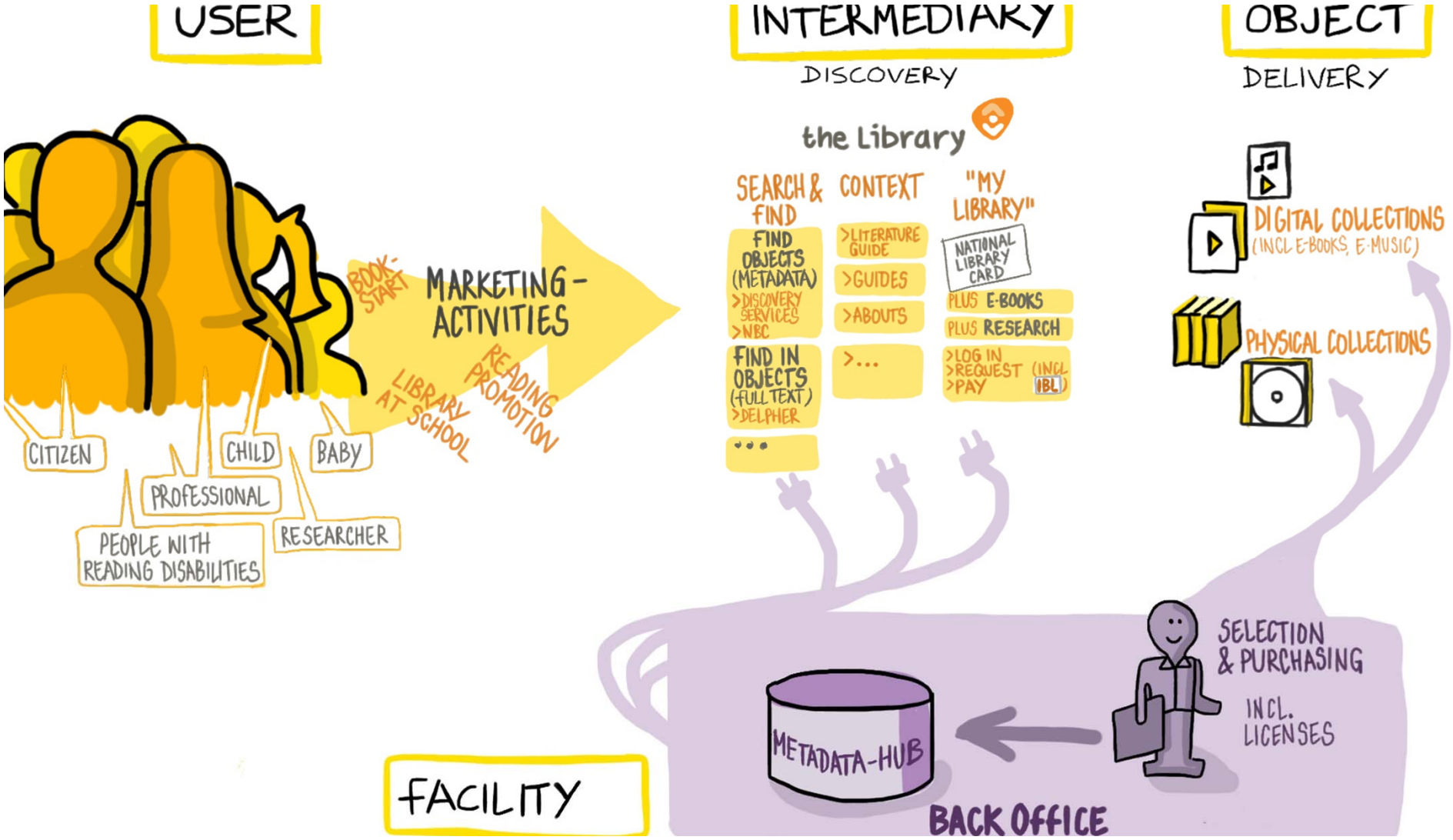
What Can University Libraries Do?

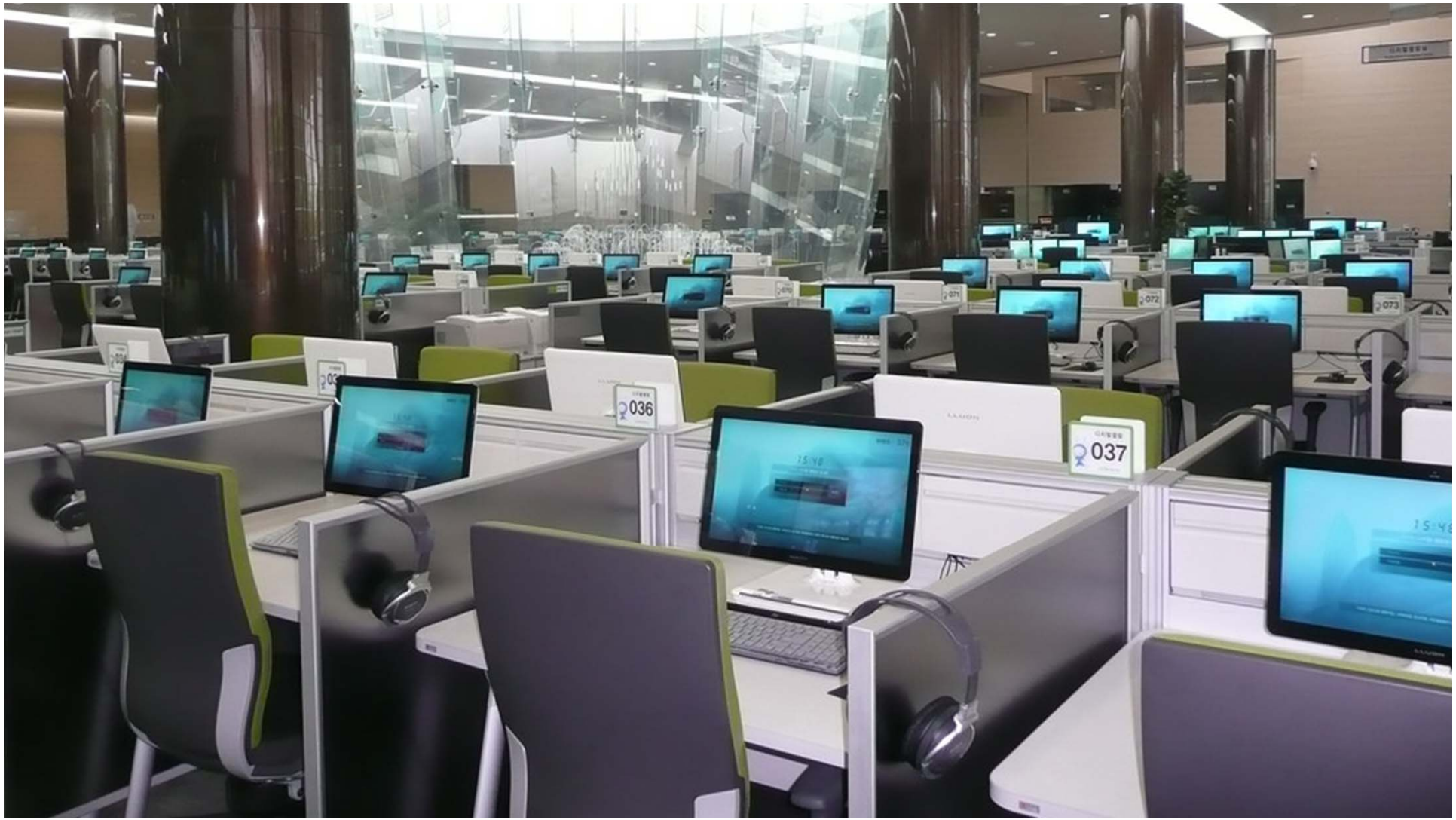
Learn more about data curation (data packaging and handling)

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


USER
INTERMEDIARY
DISCOVERY
OBJECT
DELIVERY
the Library
SEARCH & FIND
CONTEXT
"MY LIBRARY"
FIND OBJECTS (METADATA)
>DISCOVERY SERVICES
>NBC
FIND IN OBJECTS (-FULL TEXT)
>DELPHER
>LITERATURE GUIDE
>GUIDES
>ABOUTS
>...
NATIONAL LIBRARY CARD
PLUS E-BOOKS
PLUS RESEARCH
>LOG IN
>REQUEST (INCL IBL)
>PAY
DIGITAL COLLECTIONS
(INCL E-BOOKS, E-MUSIC)
PHYSICAL COLLECTIONS
BOOK START
MARKETING-ACTIVITIES
READING PROMOTION
LIBRARY AT SCHOOL
CITIZEN
CHILD
BABY
PROFESSIONAL
PEOPLE WITH READING DISABILITIES
RESEARCHER
SELECTION & PURCHASING
INCL. LICENSES
METADATA-HUB
BACK OFFICE
FACILITY


036
037


Data Handling
DATA
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

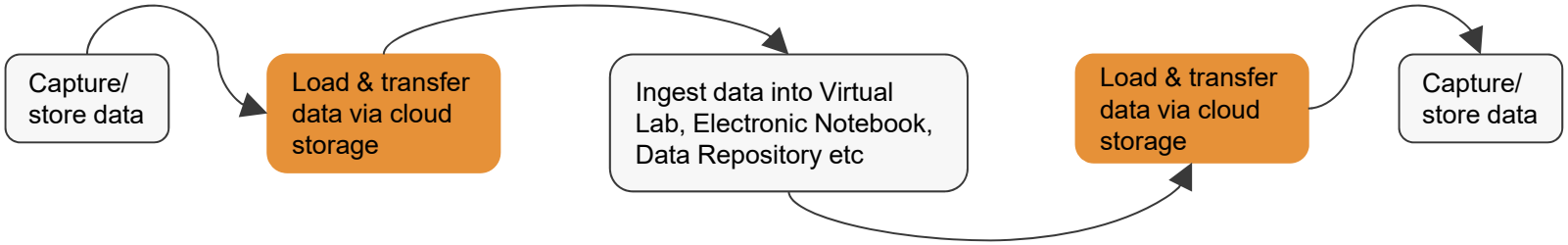
Capacity to:

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

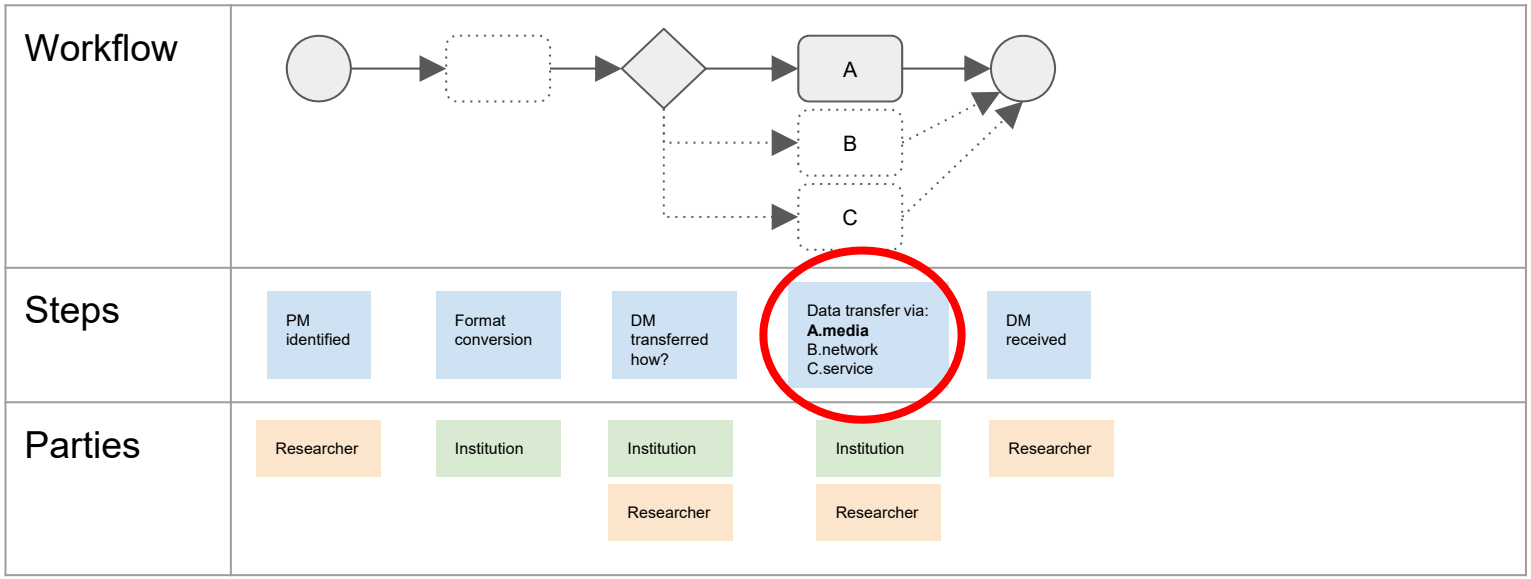
Skills, tools and services to:

- **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

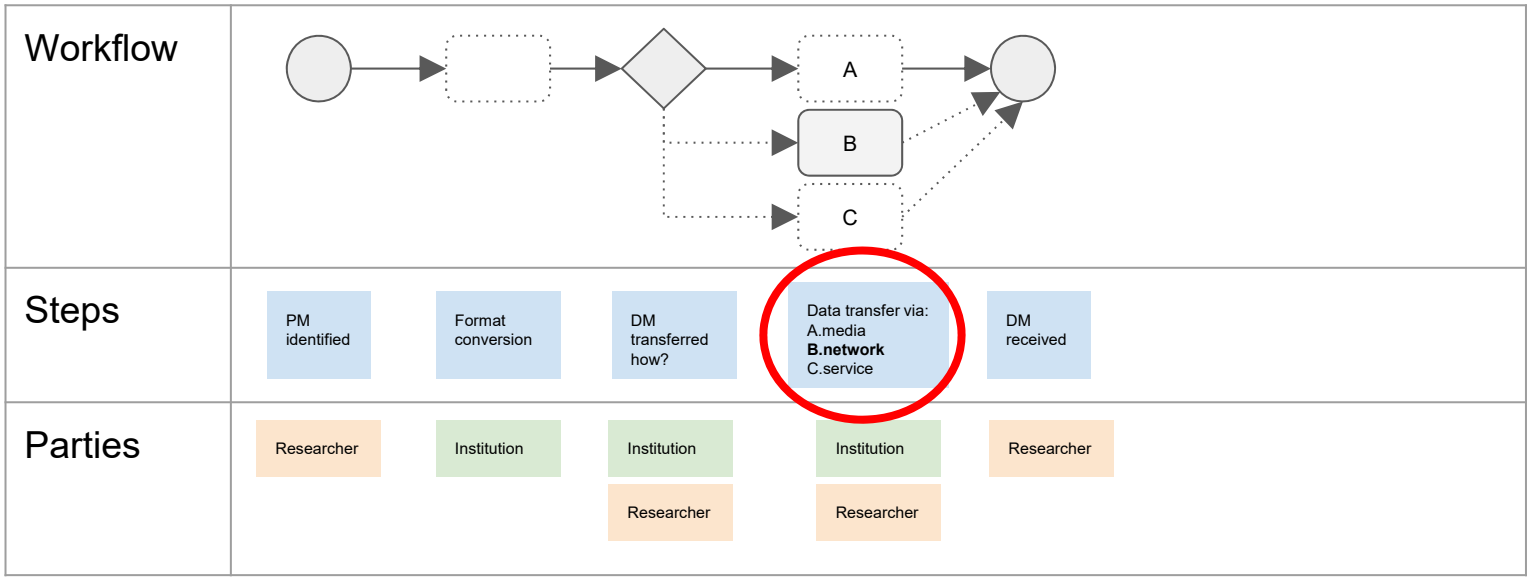


Format Conversion - InstitutionA > University



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?

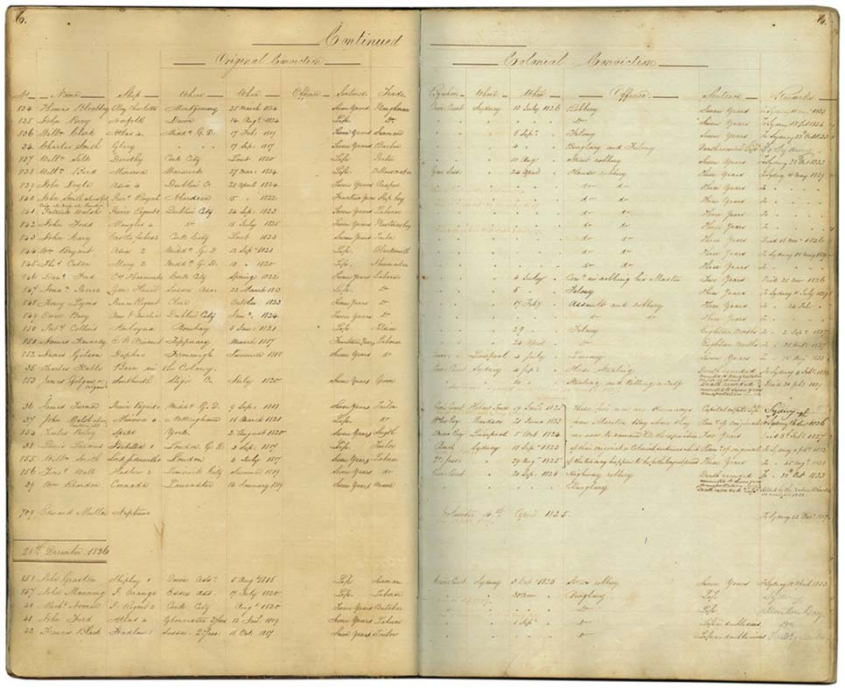
How “chunky” or “big” cultural data may be transferred and access to it enabled by the university library?

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

aarnet

Scenario 1

A CRIMINAL Life
Clue: There's no escape without a ticket.



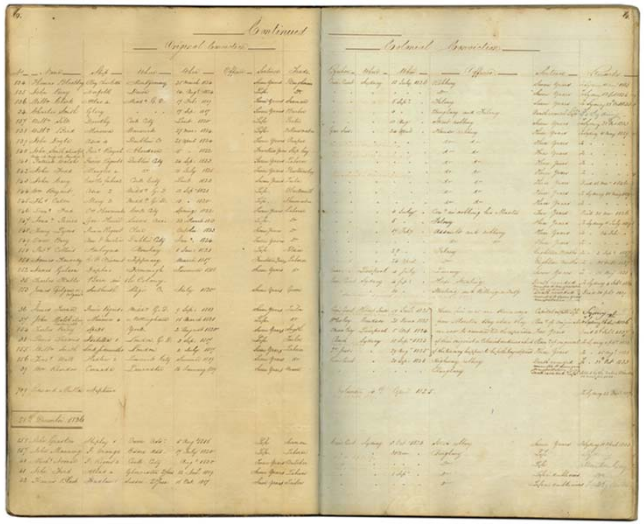
Scenario 2

CITY SPIRIT
Clue: Mind your language!

What YOU Do?


?
1.
confirmation email
DB
machine interpetation
email notification
librarians
2.
Advanced

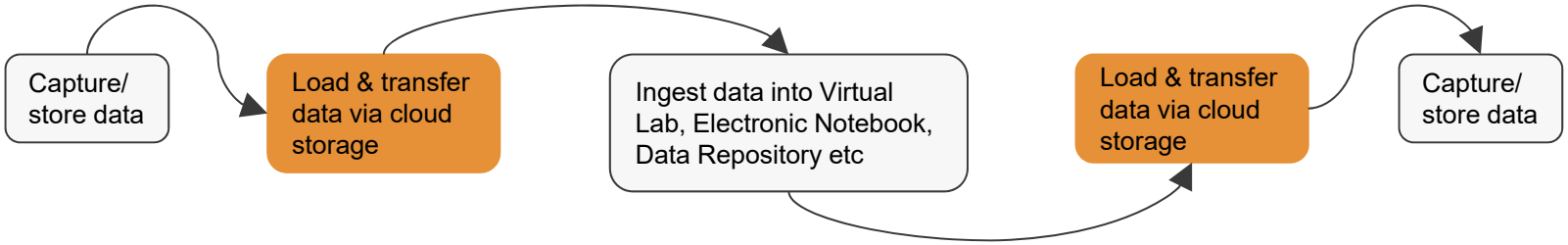
Case Studies (based on real examples)



Scenario 1: A Criminal Life

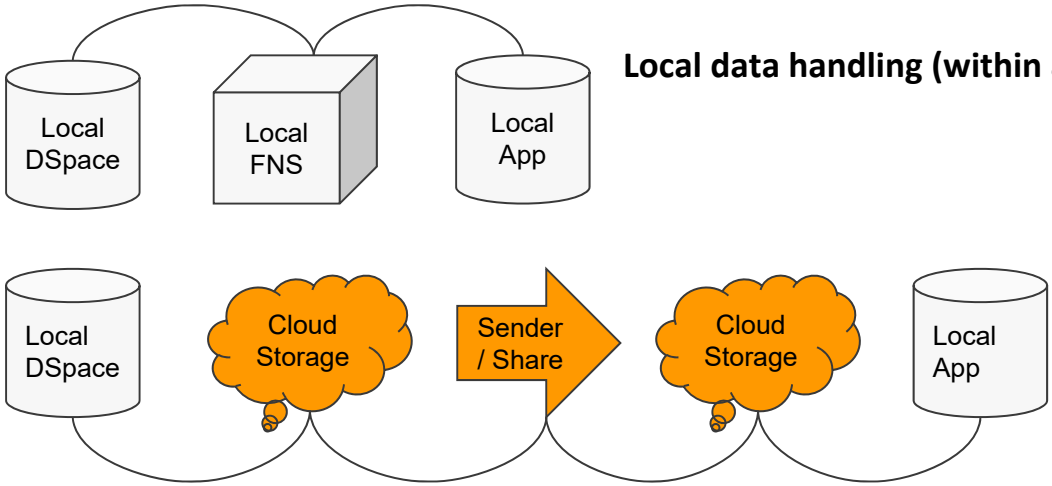
Scenario 2: City Spirit

Data Handling Workflow



Data Handling Workflow

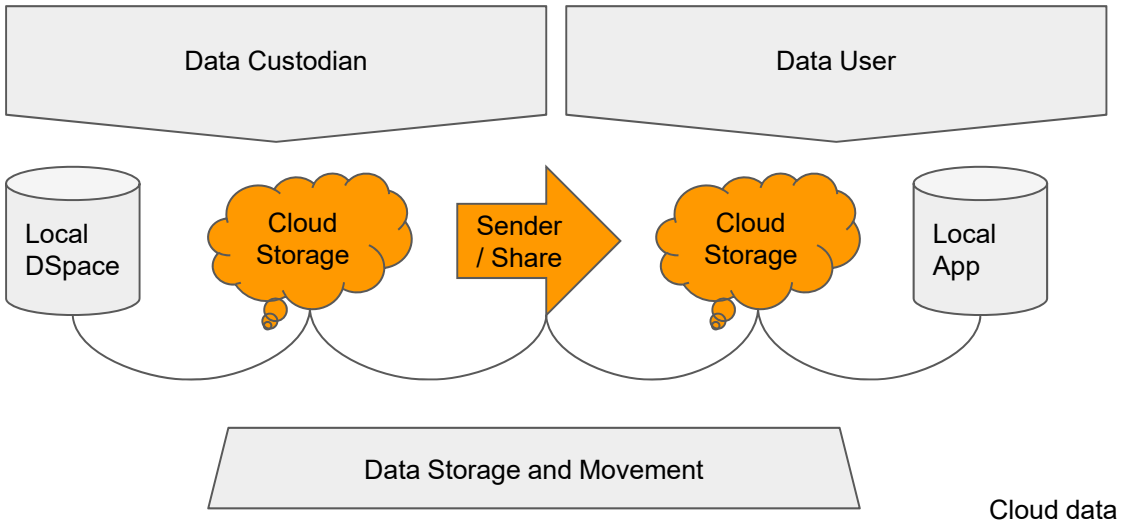
Digital assets already in digital repository hosted locally or remotely.



Local data handling (within an organisation)

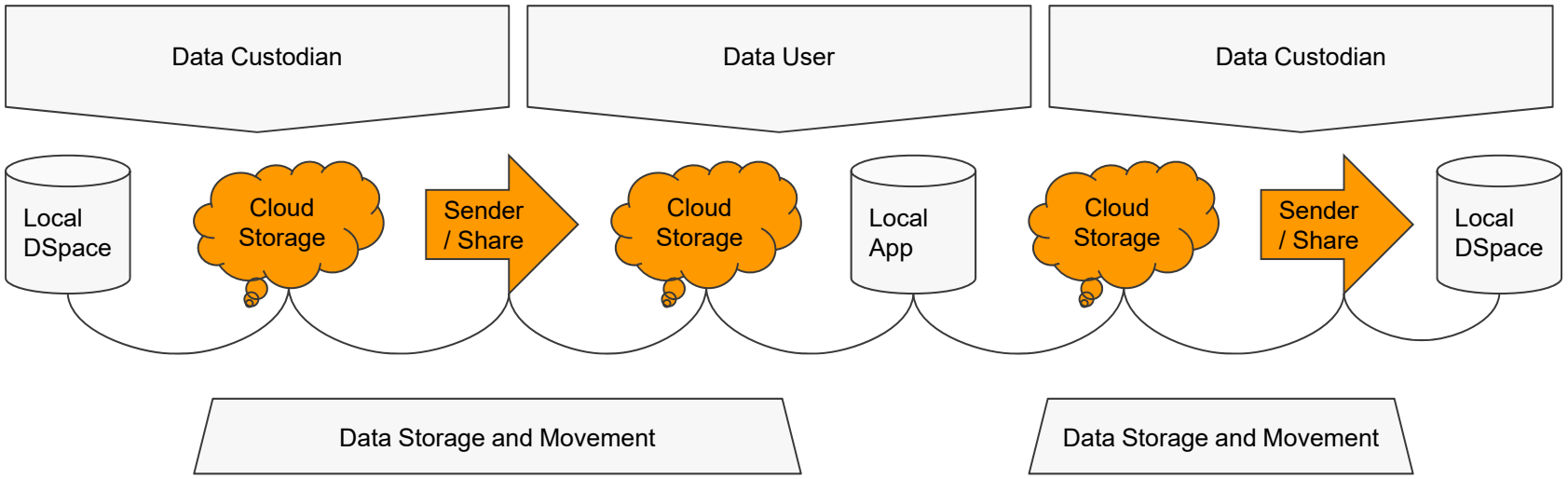
Shared data handling (within or across organisations)

Shared Data Handling

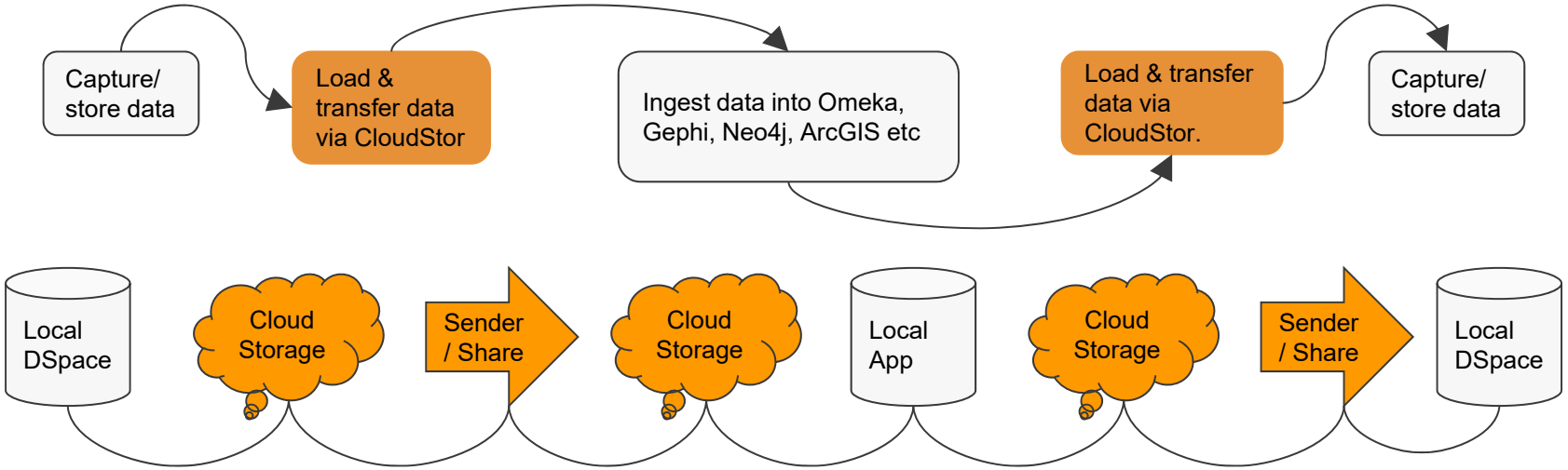


Cloud data storage and movement tools
CloudStor, Local, OneDrive, Google Drive, DropBox

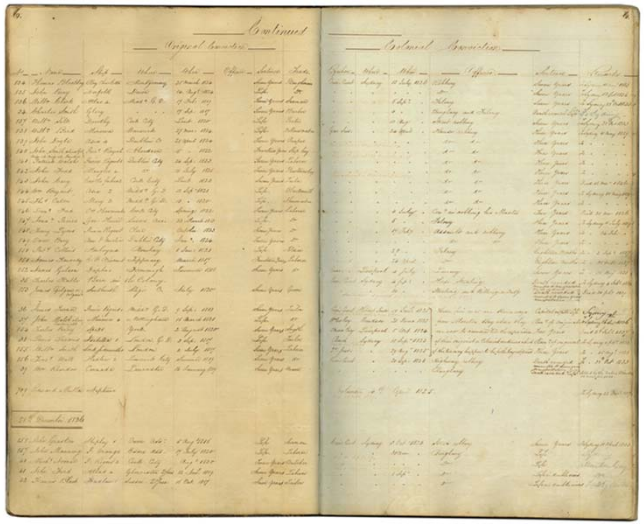
Data Movement Within/Across Organisations



End-to-End Process



Case Studies (based on real examples)

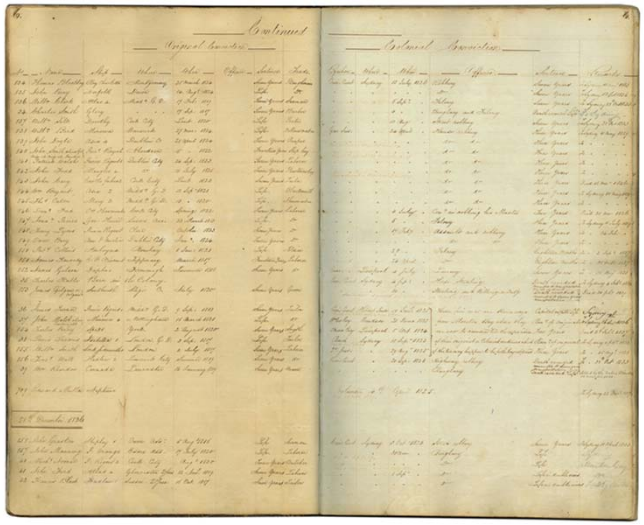


[Scenario 1: A Criminal Life]

[Scenario 2: City Spirit]

The Big Reveal...

Case Studies (the real ones)



Scenario 1: Prosecution Project

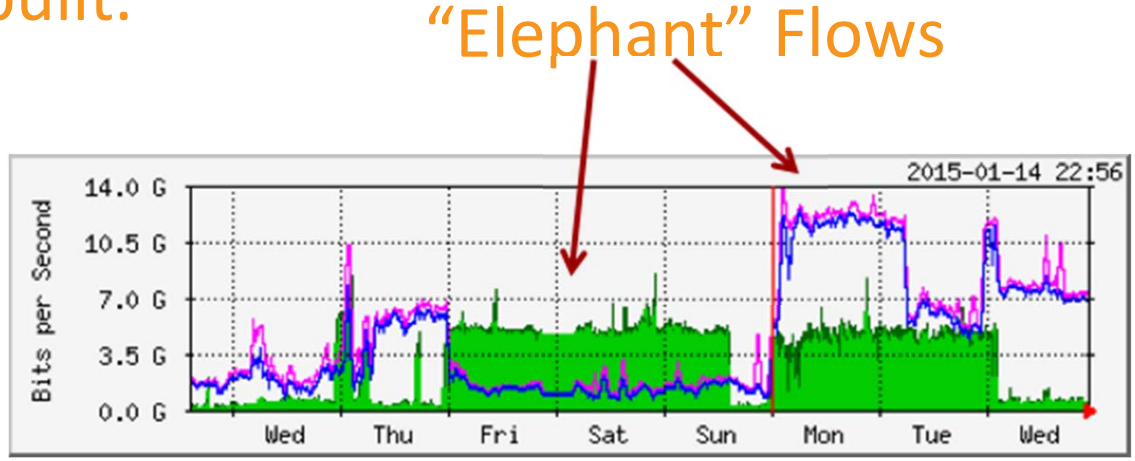
Scenario 2: QuakeBox Project

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.



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**

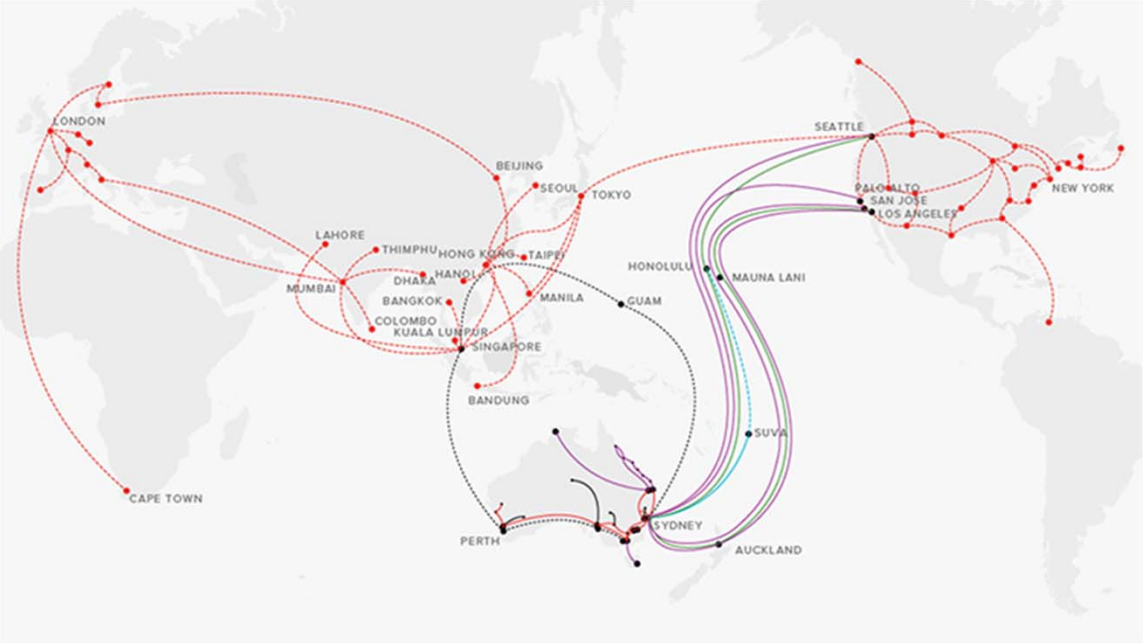
NRENs

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**

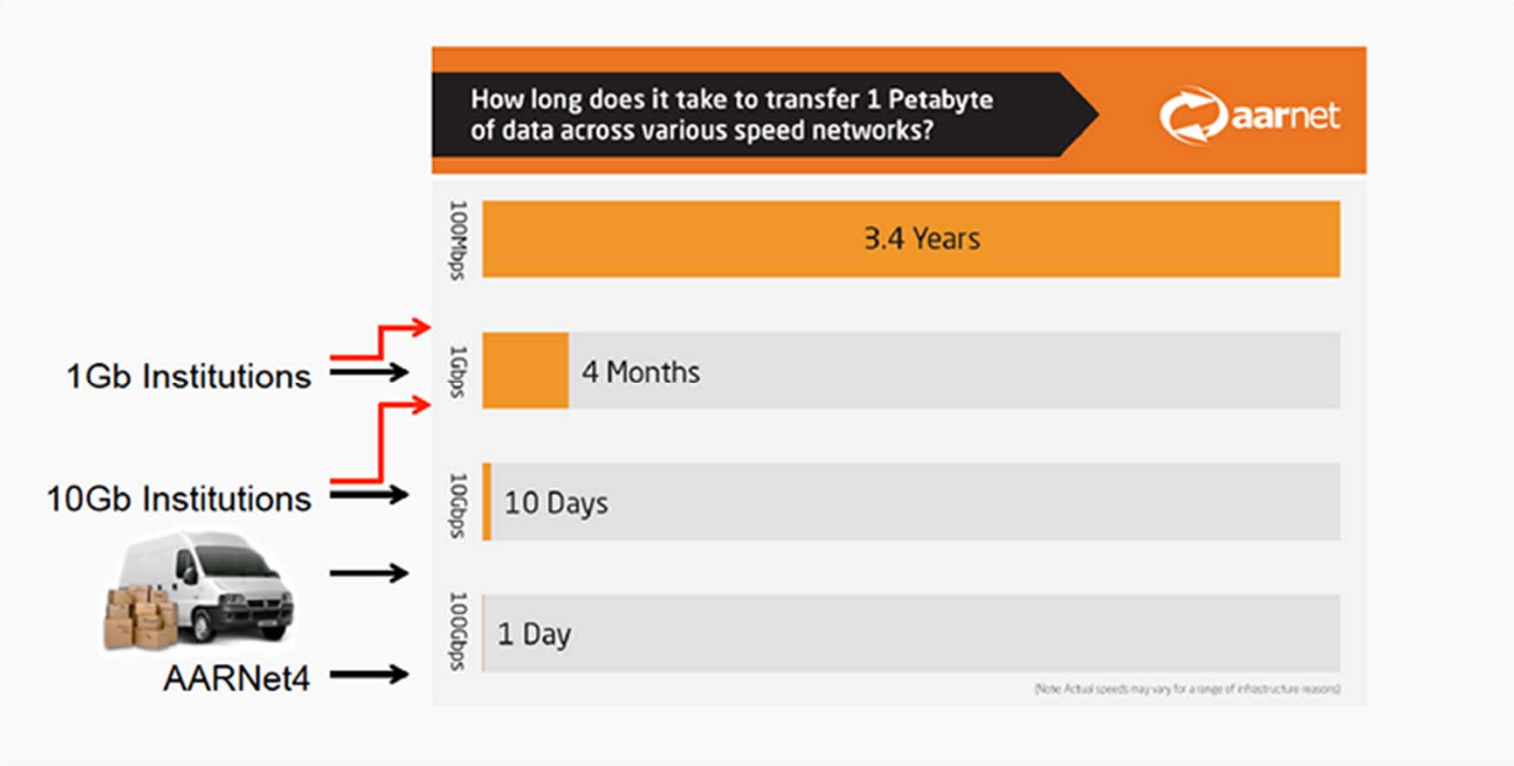
NRENs


LONDON
CAPE TOWN
BEIJING
SEOUL
TOKYO
SEATTLE
PALO ALTO
SAN JOSE
LOS ANGELES
NEW YORK
LAHORE
THIMPHU
HONG KONG
TAIPEI
HANOI
DHAKA
MUMBAI
BANGKOK
MANILA
GUAM
HONOLULU
MAUNA LANI
COLOMBO
KUALA LUMPUR
SINGAPORE
BANDUNG
SUVA
PERTH
SYDNEY
AUCKLAND

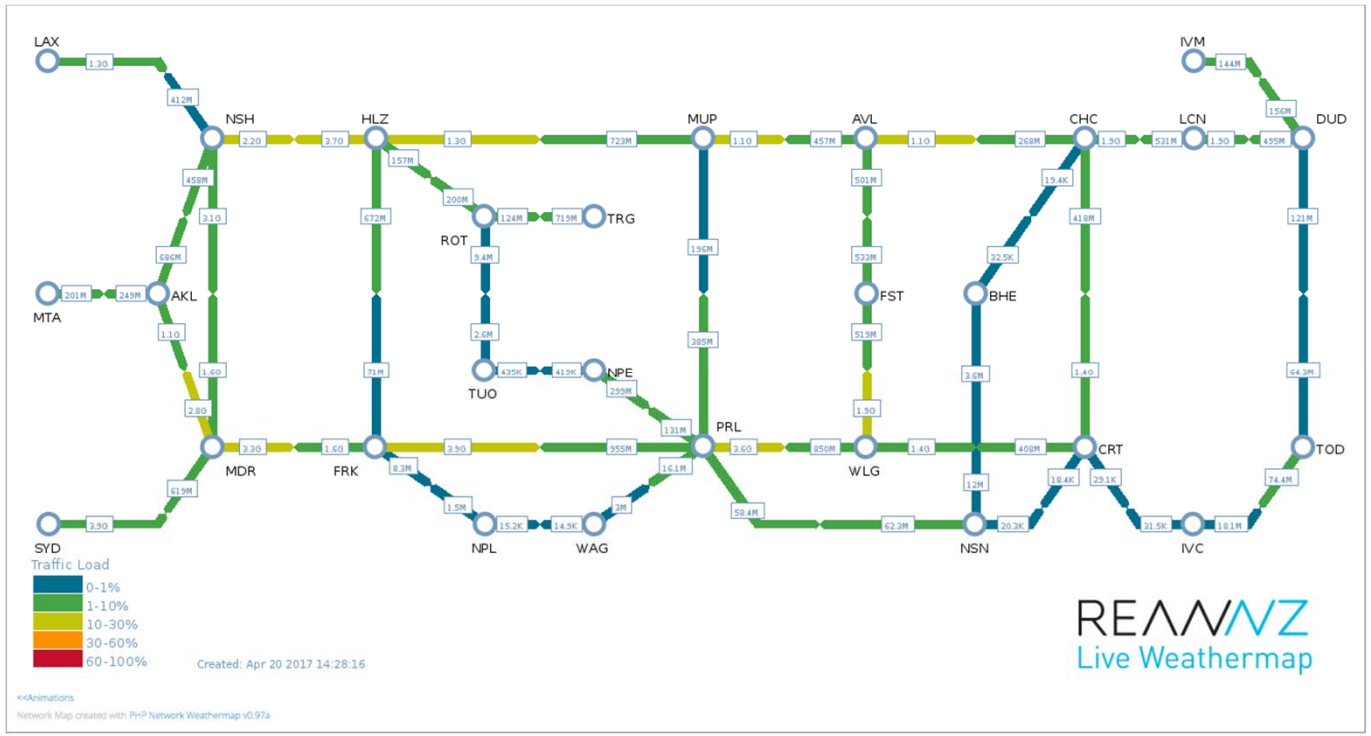
Australia's National Broadband Network

Download (Mbps)	Upload (Mbps)
12	1
25	5
25	10
50	20
100	40

Australia's NREN



New Zealand's NREN



aarnet

Data Mining in the Sky (In the Field Stories)

“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**

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

100 terabytes (100TB) = **~119357 h 38 m**

Download speed: **~3.38 Mbps**

(Upload speed: 0.50 Mbps)

Speed Test - Downloading at a Conference

100 megabytes (100MB) = **~17 s**

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

100 terabytes (100TB) = **~4793 h 1 m**

Download speed: **~54 Mbps**

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.**

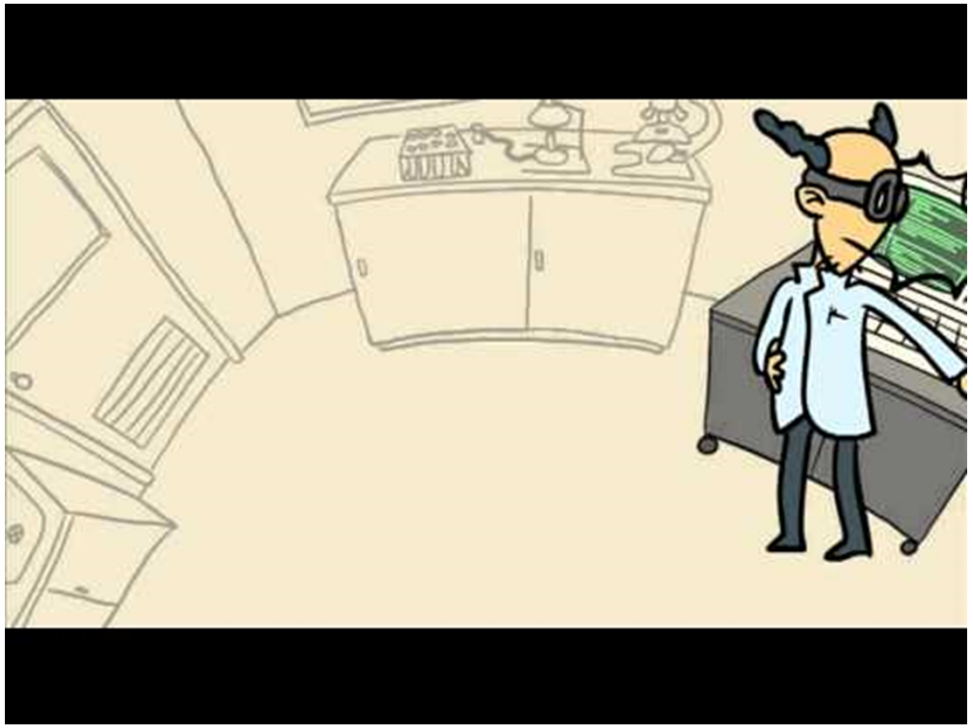
Data Access and Transfer Tools

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

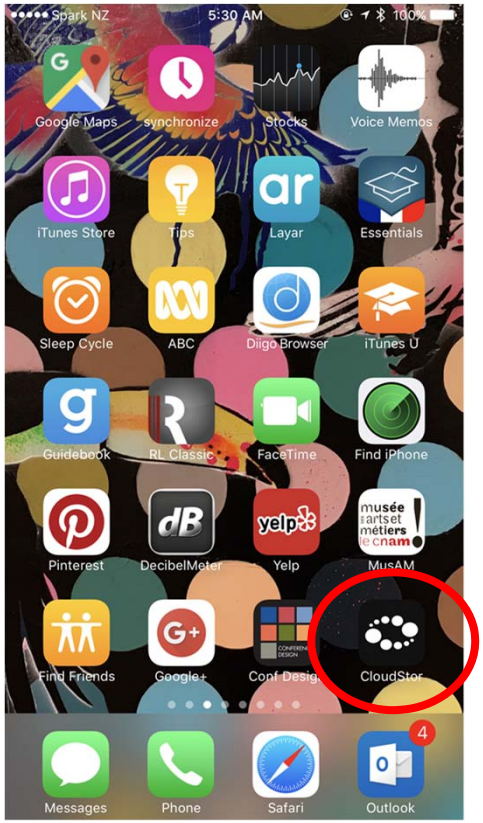
Make data transferrable (conditionally) e.g. data transferred (from a A to B) is defined a (one off, recurrent, syncing)

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

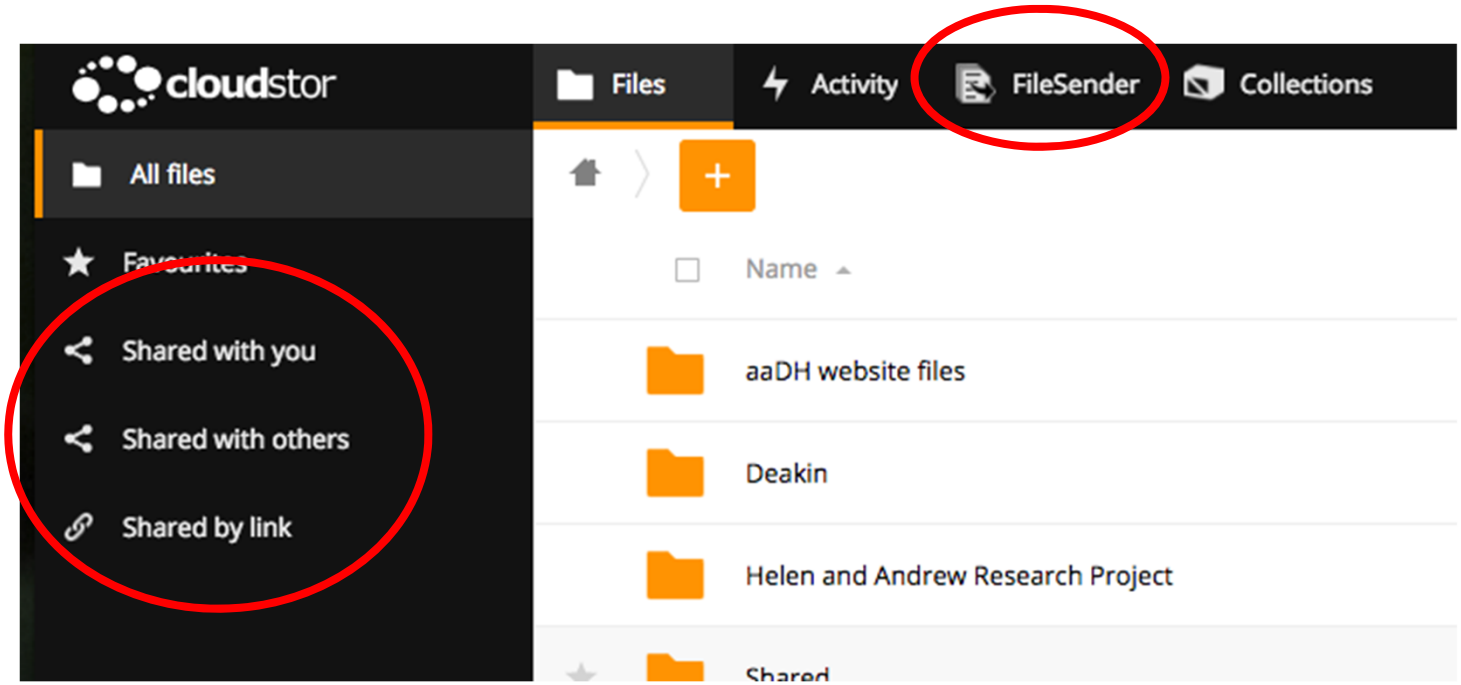
CloudStor



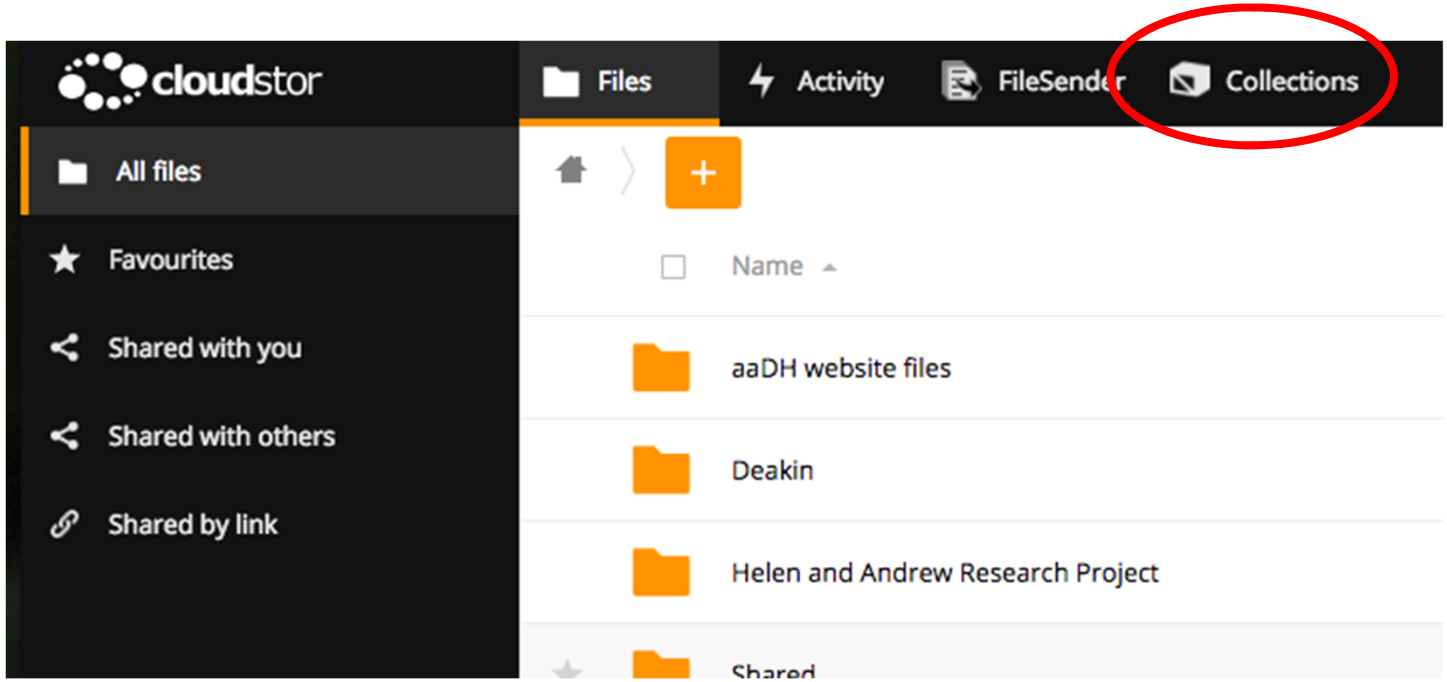
Synching


Spark NZ
5:30 AM
100%
Google Maps
synchronize
Stocks
Voice Memos
iTunes Store
Tips
Layar
Essentials
Sleep Cycle
ABC
Diigo Browser
iTunes U
Guidebook
RL Classic
FaceTime
Find iPhone
Pinterest
DecibelMeter
Yelp
MusAM
Find Friends
Google+
Conf Desig
CloudStor
Messages
Phone
Safari
Outlook
4

Sharing


cloudstor
Files
Activity
FileSender
Collections
All files
Shared with you
Shared with others
Shared by link
Name
aaDH website files
Deakin
Helen and Andrew Research Project
Shared

CloudStor Collections Plugin



CloudStor Collections Plugin

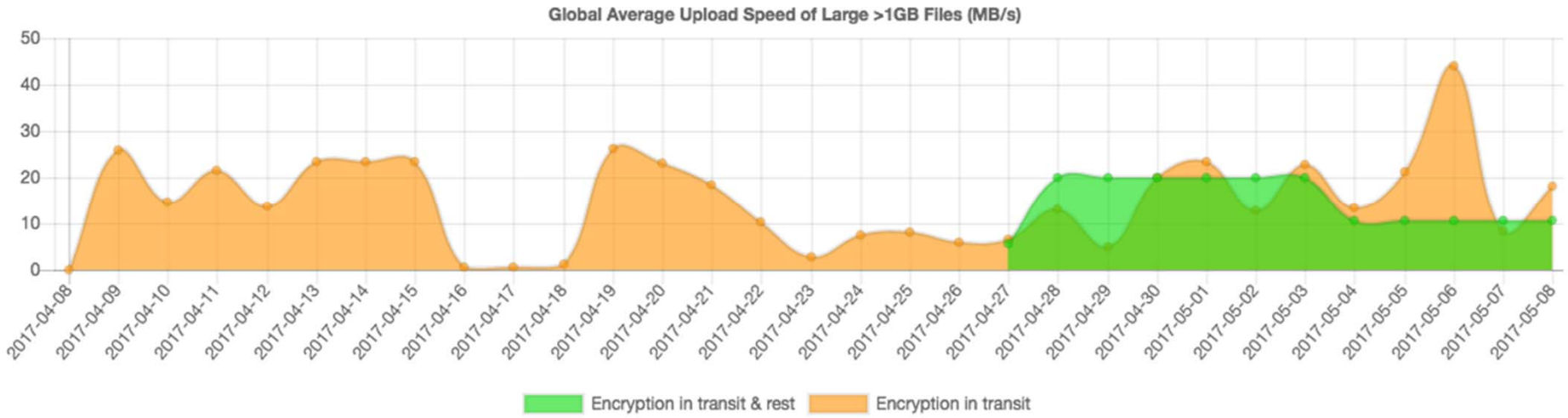
Enables researchers to easily collate, annotate, package and share groups of files and metadata, and is designed to support research data packaging for the long tail of research groups (MB-TB-GB scale)

Features of v1.1

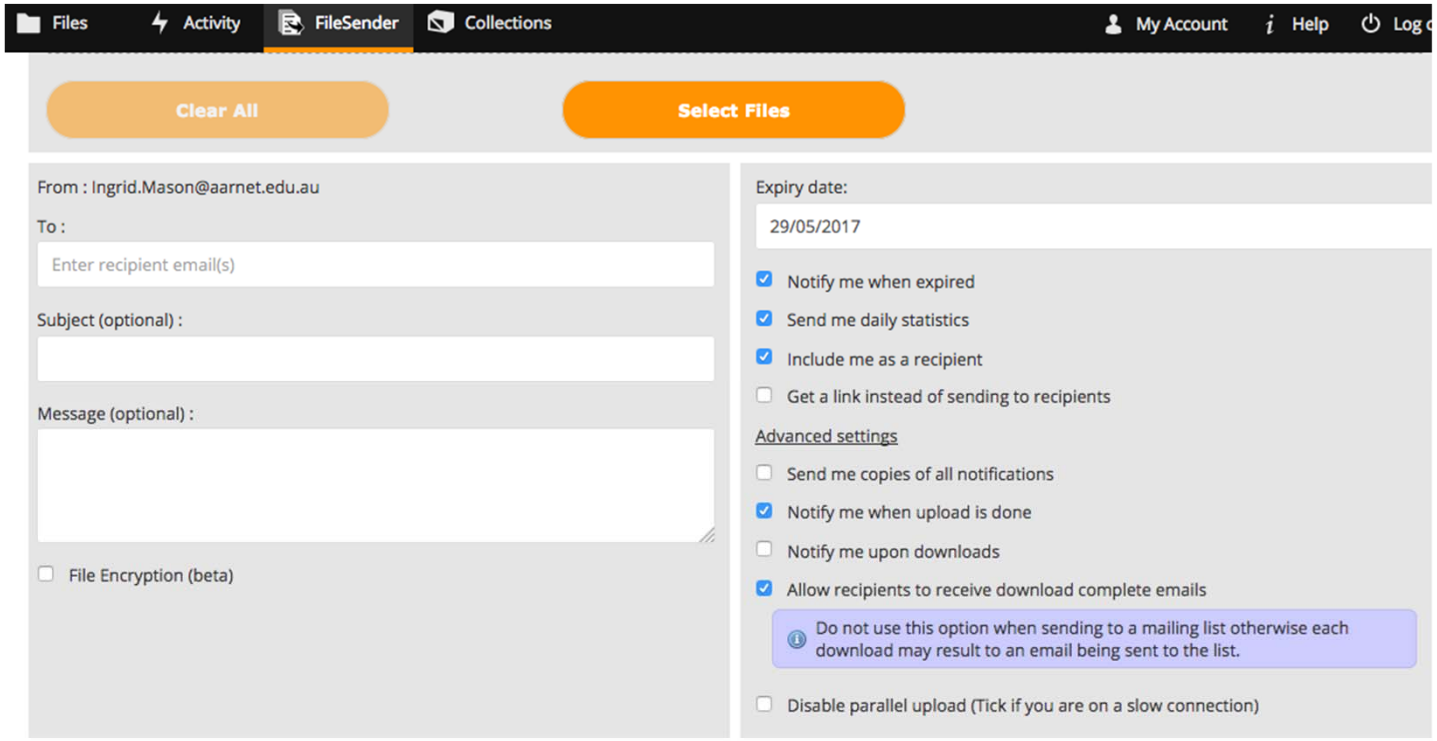
- **Select any destination for data package transfer.**
- **Readme file (RDFa) machine+human readable options.**
- **Select different metadata model for different ingest targets.**

aarnet

Encryption (at rest & in transit)



Encryption (at rest & in transit)



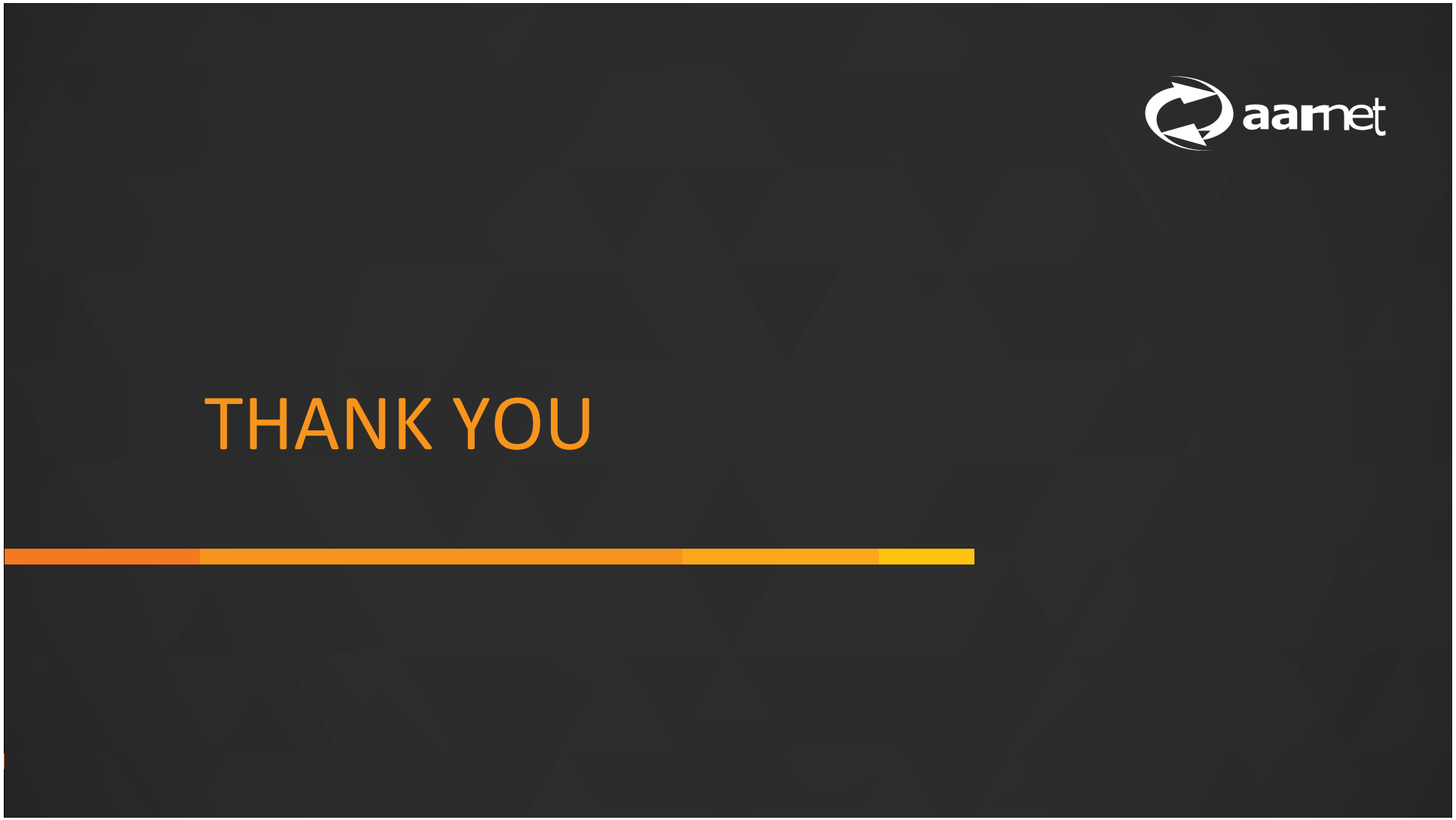
Next Steps

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).


aarnet
THANK YOU

Case Study - Prosecution Project along with the Cultures and Community Project (Open API)

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

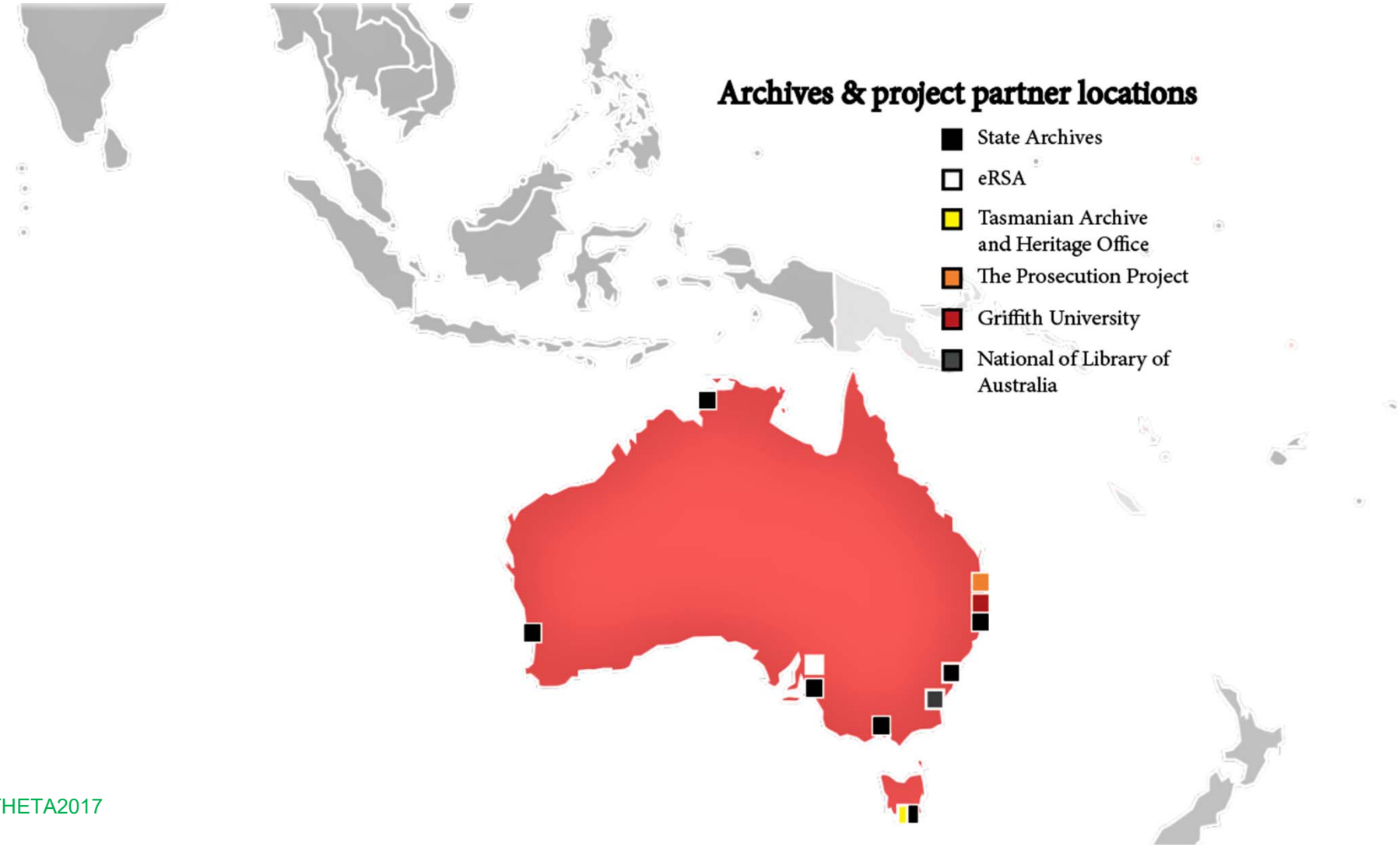
Griffith UNIVERSITY

#THETA2017

Liaison with cultural institutions for collection access

Griffith UNIVERSITY


The barriers
Negotiations required
Old Bailey
Legacy Systems
Griffith
UNIVERSITY


Archives & project partner locations
State Archives
eRSA
Tasmanian Archive and Heritage Office
The Prosecution Project
Griffith University
National of Library of Australia

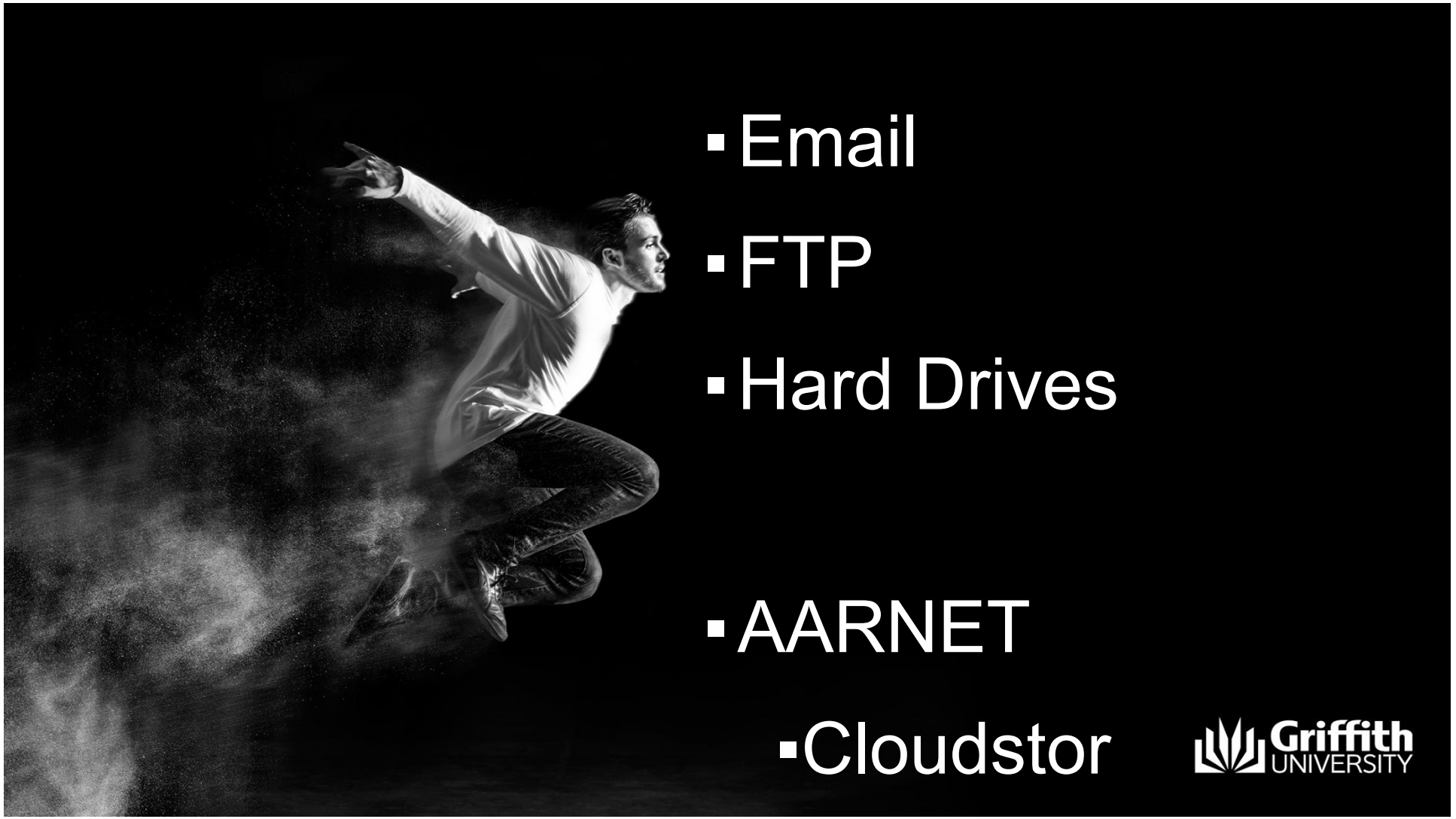
#THETA2017

Reuse of digitised material and arrangement for digitisation

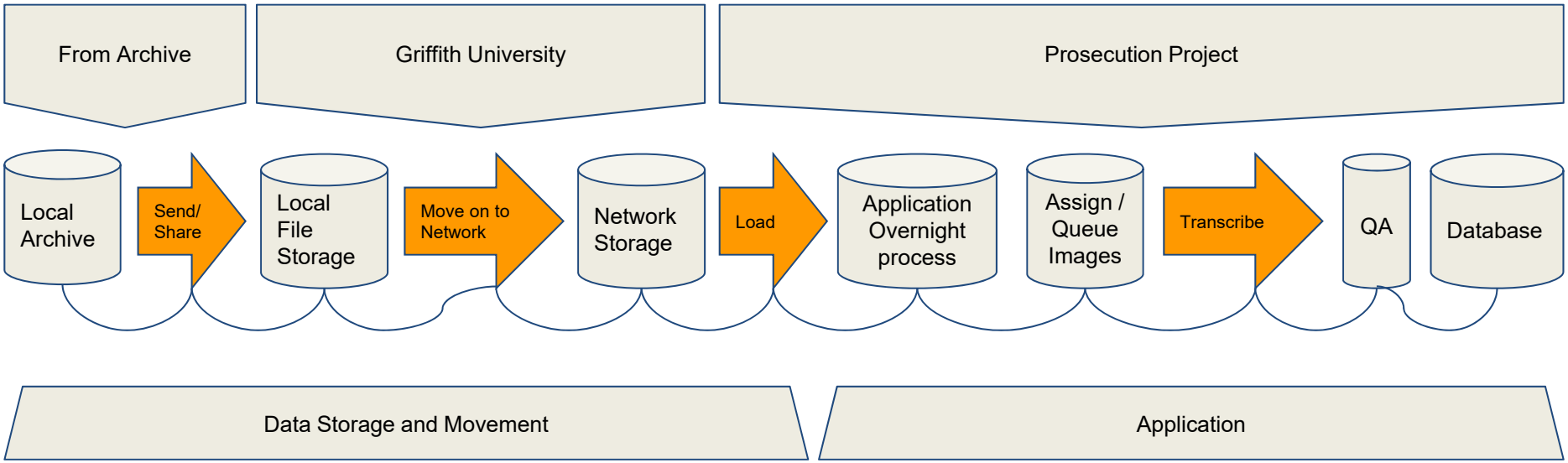
Griffith UNIVERSITY


Different Priorities
Record storage types
Records Lost
Pressures
Griffith
UNIVERSITY

Transfer of digitised material (different approaches)


Email
FTP
Hard Drives
AARNET
Cloudstor
Griffith
UNIVERSITY

Workflow of adding images to PP



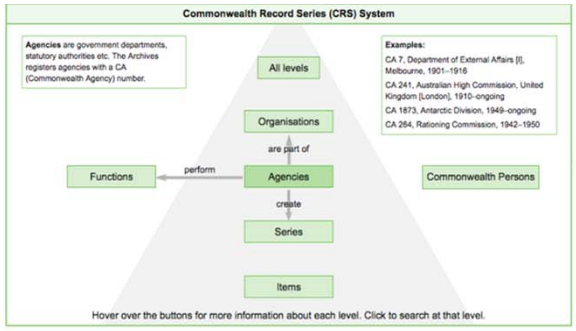
Griffith UNIVERSITY

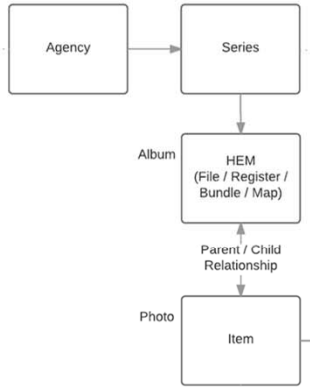
Transcription platform & 360 degree data sharing (RDS C&C)

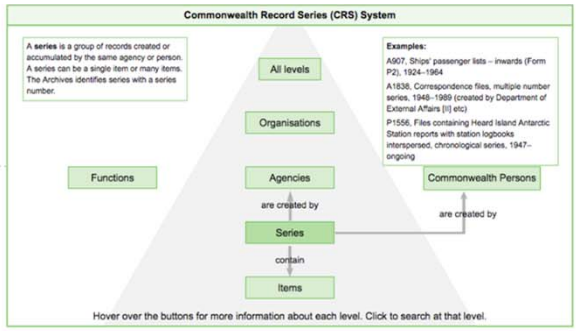
#THETA2017

Griffith UNIVERSITY

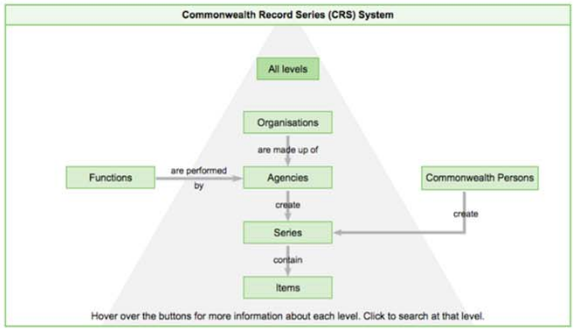

Metadata
Archives Metadata
Advisory Board
Technical Advisory Board
Griffith UNIVERSITY

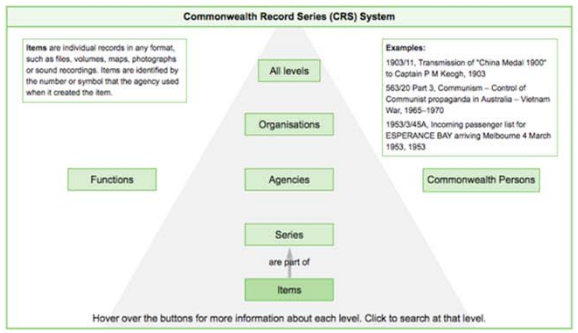






http://recordsearch.naa.gov.au/SearchNRetrieve/Interface/SearchScreens/AdvSearchMain.aspx





Griffith UNIVERSITY

Metadata Fields to Share

PP	TAHO
• Def_firname • Def_surname • offence status • trial_date • trial_id • trial_place • Verdict	• Title • Description • Publisher • Date • Type • Format • Identifier • Source • Language

Griffith UNIVERSITY

DC elements (TAHO -> PP)

Metadata Element	Data Type	Data Value
dc.title	String	Register of Indictments
dc.publisher.secondary	String	Prosecution Project
dc.publisher.primary	String	LINC Tasmania
dc.publisher.uri	URI	http://prosecutionproject.griffith.edu.au/oai/
dc.identifier.series	Identifier	AD875
dc.identifier.item	Identifier	AD875/1/1
dc.identifier.part	Identifier	AD875/1/1-1
dc.identifier.subpart	identifier	AD875/1/1-1-1
	URI	*https://linctas.ent.sirsidynix.net.au/client/en_AU/tas/search/detailnonmodal/ent:$002f$002fARCHIVES_DIGITISED$002f0$002fARCH_DIGITISED:XXXXX/one*

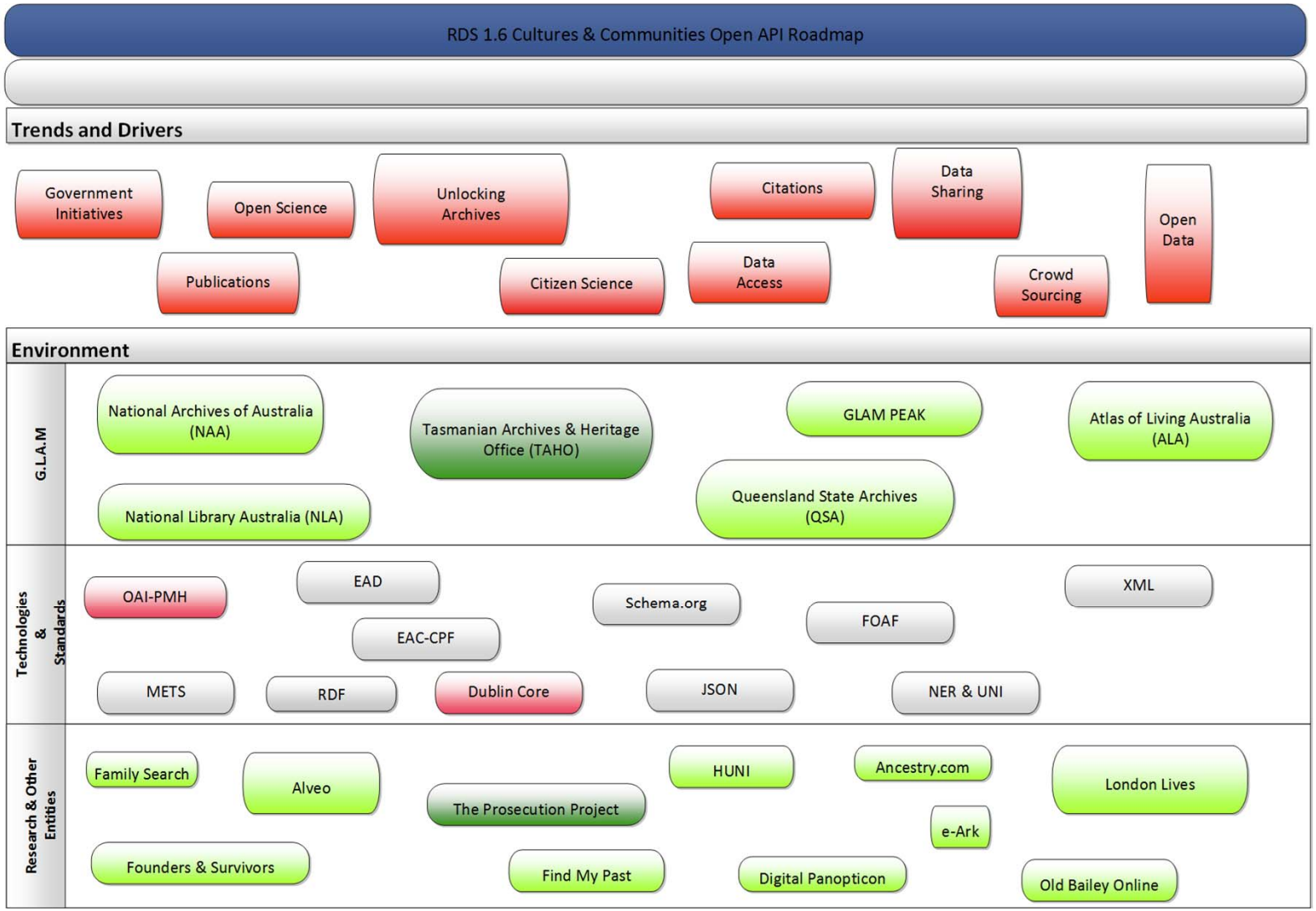
Orginal Document

Custom PP Data DC elements (PP→ TAHO)

Metadata Element	Data Type	Data Value
pp.trial.name	String	*Trial of...*
pp.trial.identifier	Identifier	*12345*
pp.trial.offence	Controlled term	*Murder*
pp.trial.date	Time/date format	
pp.trial.placename	Controlled term	
pp.trial.verdict	Controlled term	
pp.person.firstname	String	
pp.person.lastname	String	
pp.rights.licence	Controlled term	

Griffith UNIVERSITY

Orginal Document


RDS 1.6 Cultures & Communities Open API Roadmap
Trends and Drivers
Government Initiatives
Open Science
Unlocking Archives
Citations
Data Sharing
Open Data
Publications
Citizen Science
Data Access
Crowd Sourcing
Environment
G.L.A.M
National Archives of Australia (NAA)
Tasmanian Archives & Heritage Office (TAHO)
GLAM PEAK
Atlas of Living Australia (ALA)
National Library Australia (NLA)
Queensland State Archives (QSA)
Technologies & Standards
OAI-PMH
EAD
Schema.org
XML
FOAF
EAC-CPF
METS
RDF
Dublin Core
JSON
NER & UNI
Research & Other Entities
Family Search
Alveo
HUNI
Ancestry.com
London Lives
The Prosecution Project
e-Ark
Founders & Survivors
Find My Past
Digital Panopticon
Old Bailey Online

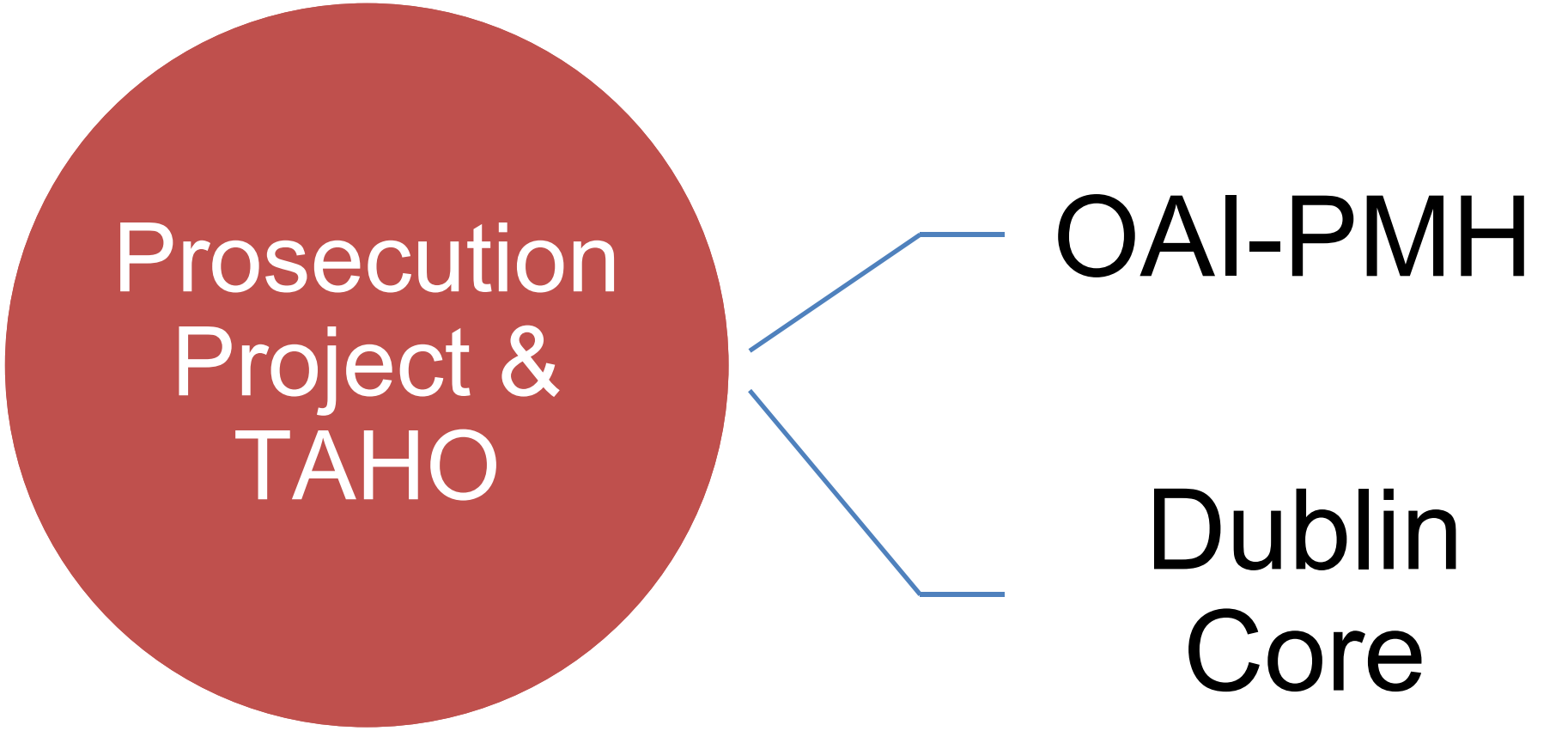
Options: Standards – Transport Layers

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

Griffith UNIVERSITY

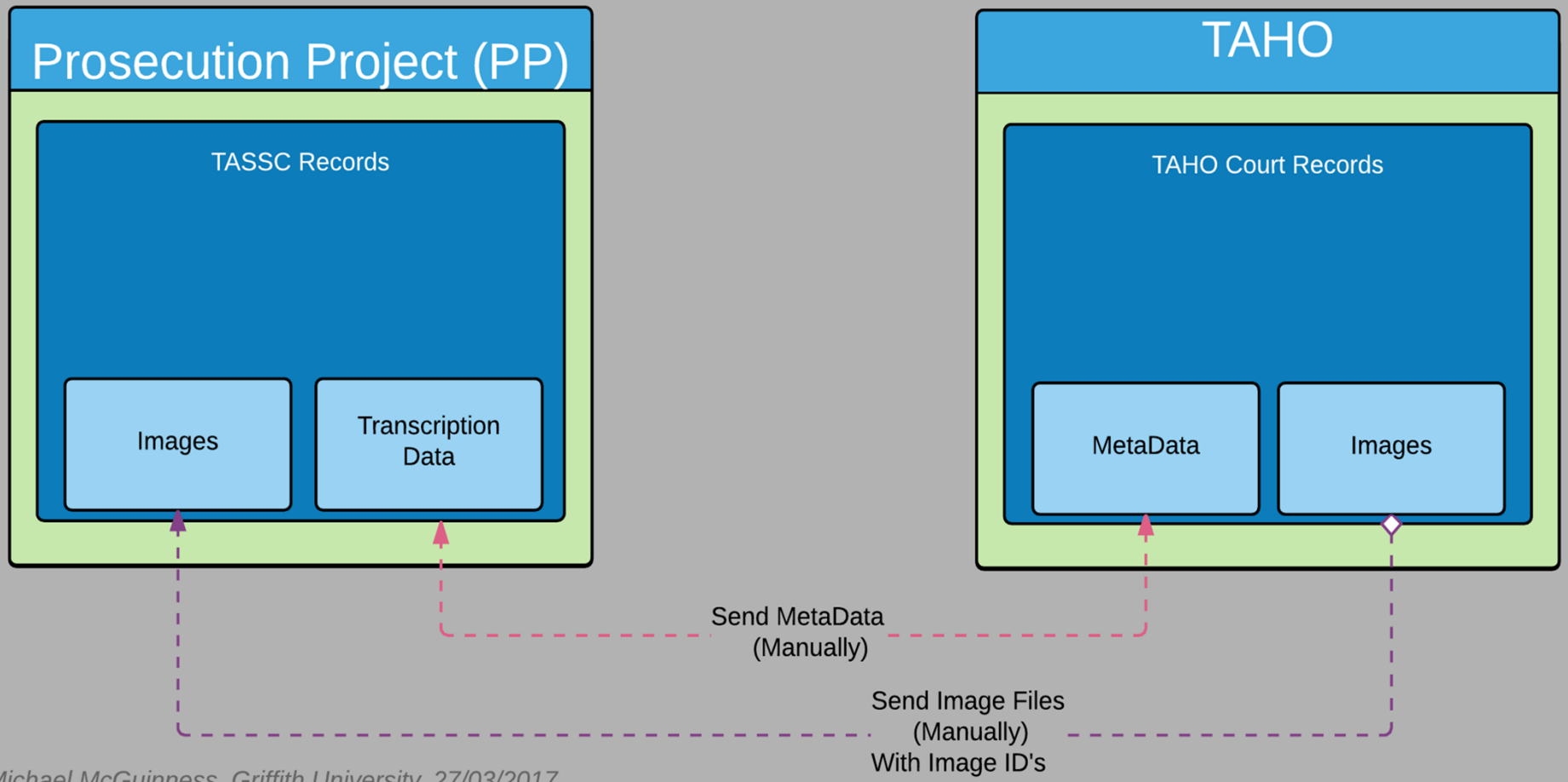
Proposed Solution/Pilot

Griffith UNIVERSITY



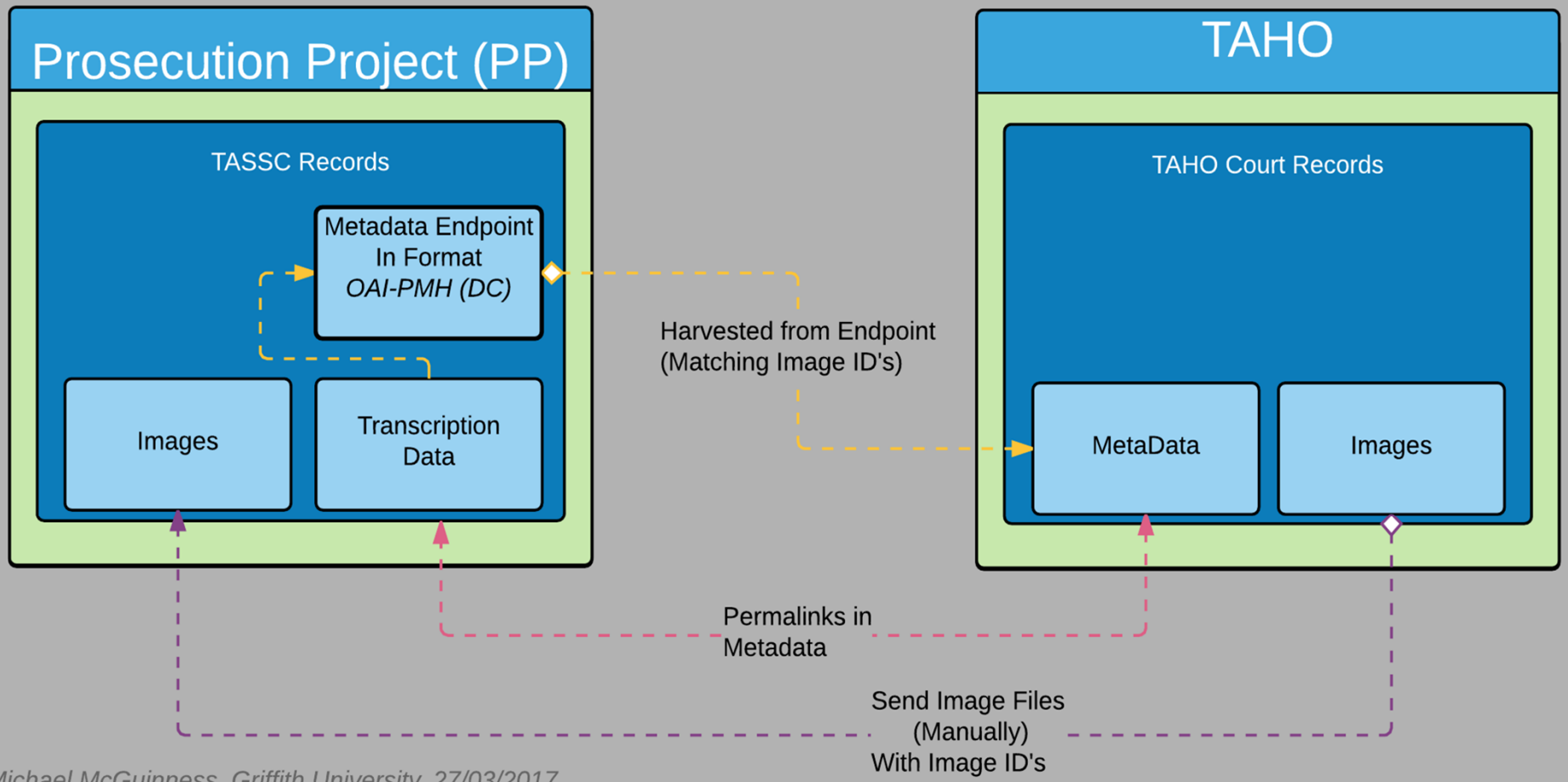
Griffith UNIVERSITY

Current State



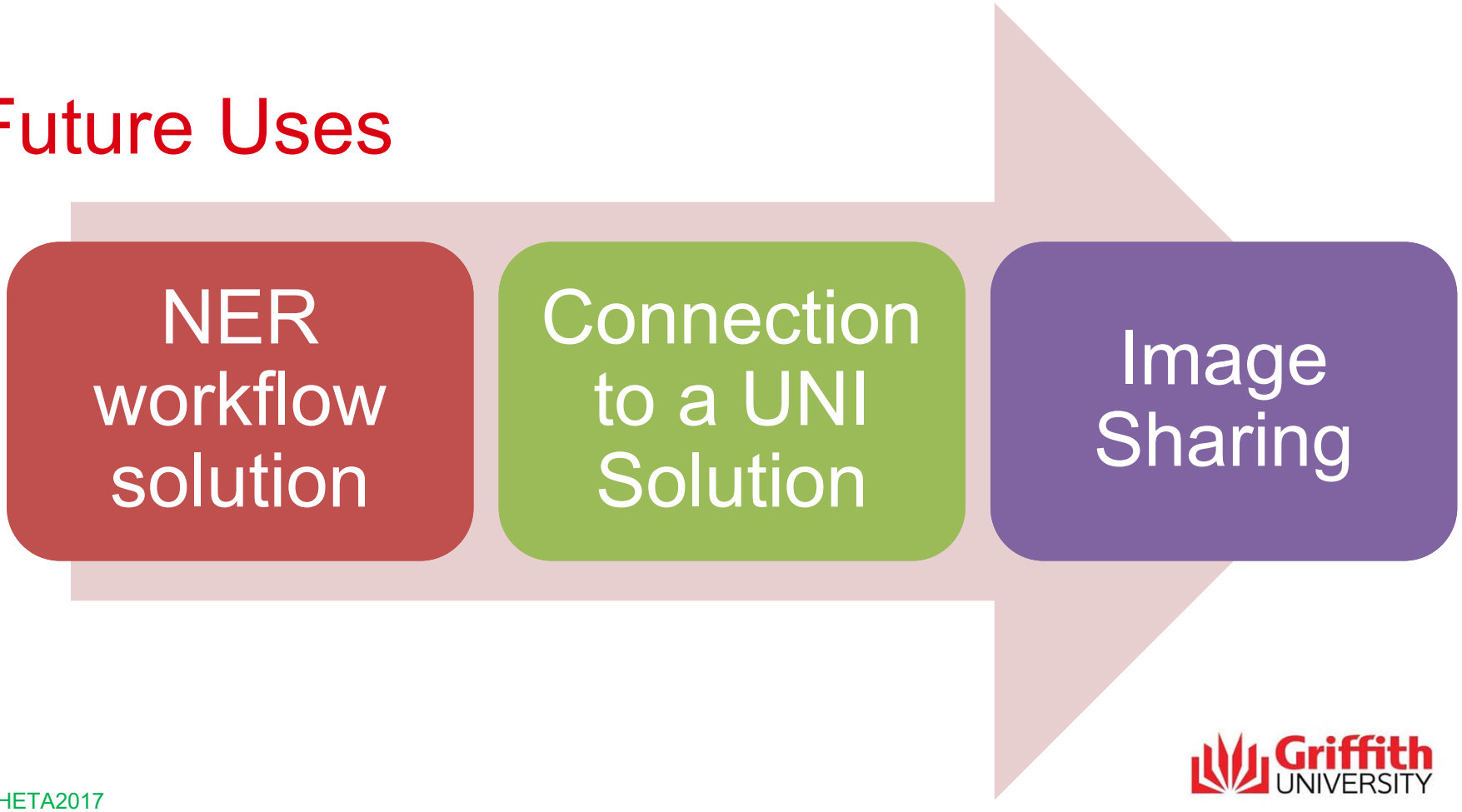
Michael McGuinness, Griffith University, 27/03/2017

Next State



Michael McGuinness, Griffith University, 27/03/2017

Future Uses



Griffith UNIVERSITY

Poster Reveal

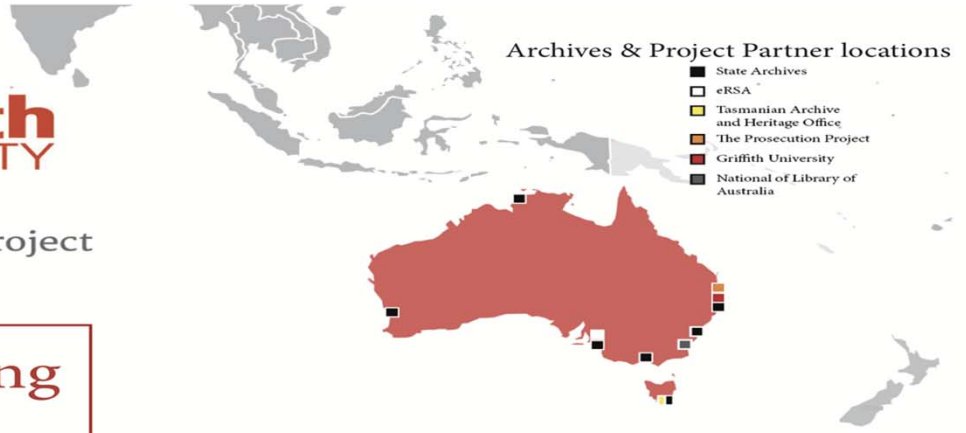
#THETA2017

Griffith UNIVERSITY

eRSA
Advancing Research Innovation

Griffith UNIVERSITY

The Prosecution Project



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?

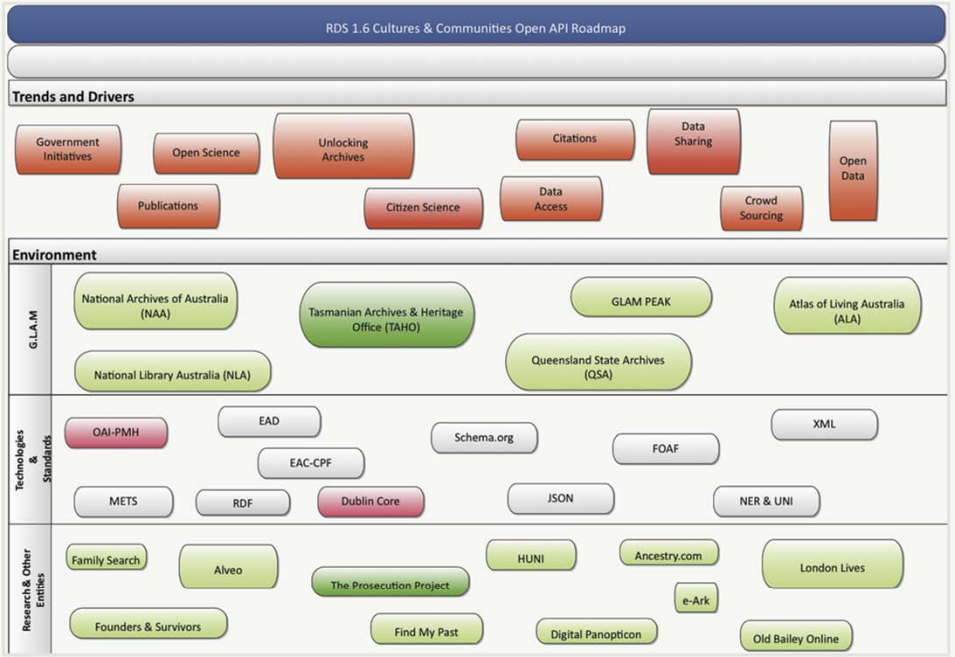
Alongside stakeholders, such as National Library of Australia, National Australian Archives, Tasmanian Archives and Heritage Office (TAHO) and Queensland State Archives, we're working on the Open API Project through a program of work managed by eRSA.

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.

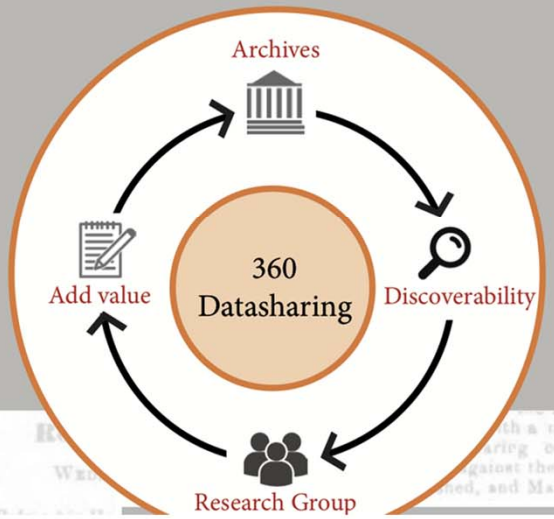


www.ersa.edu.au/cultures-community-project | www.prosecutionproject.griffith.edu.au | comms@ersa.edu.au

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

Griffith UNIVERSITY


THANK YOU
Questions?
Griffith
UNIVERSITY

Workshop Summary

#THETA2017

- 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)