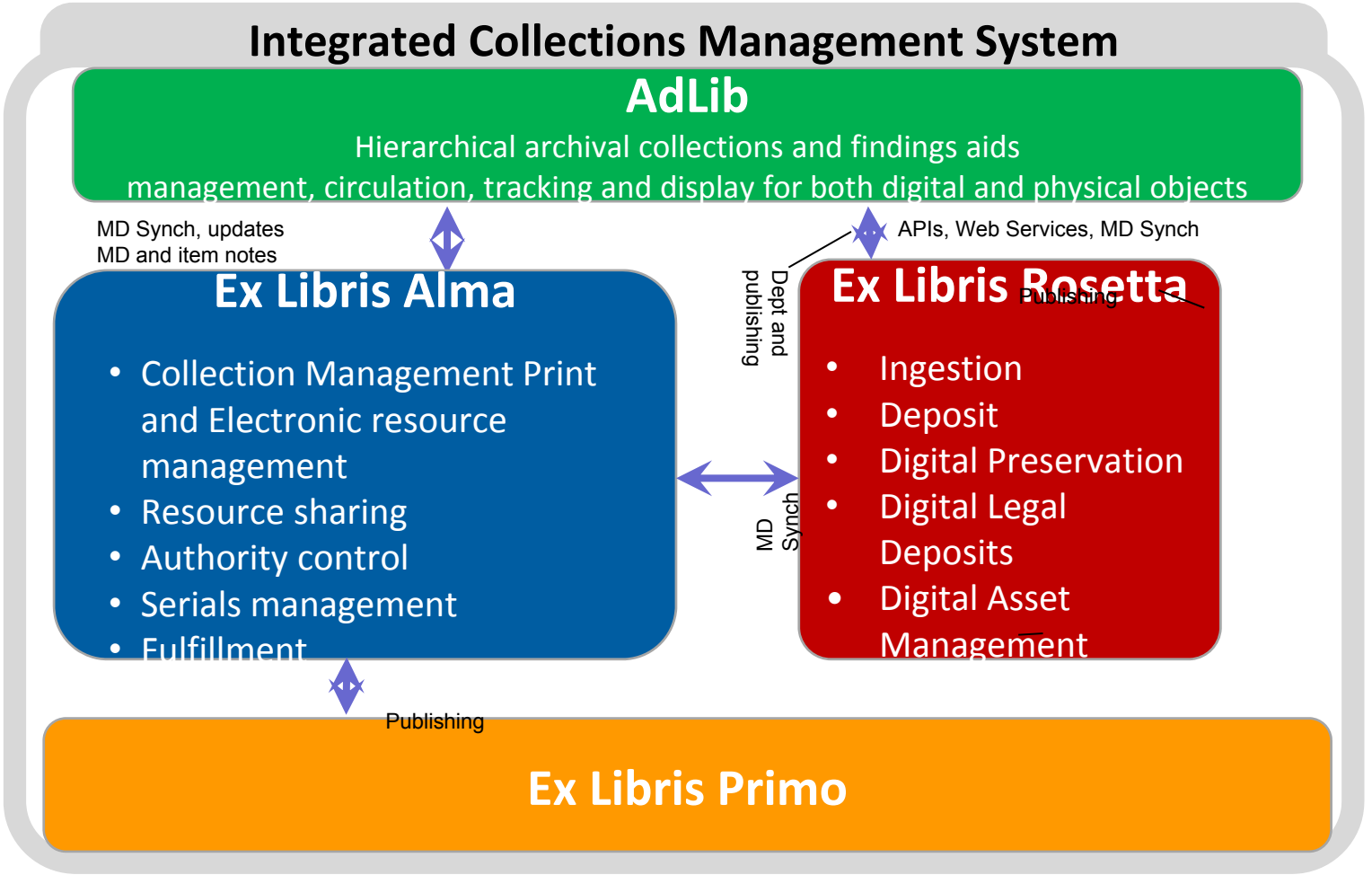
Mass Digitisation for Mass Preservation
At
State Library of New South Wales

Digital Infrastructure Program - part of Digital Excellence Program

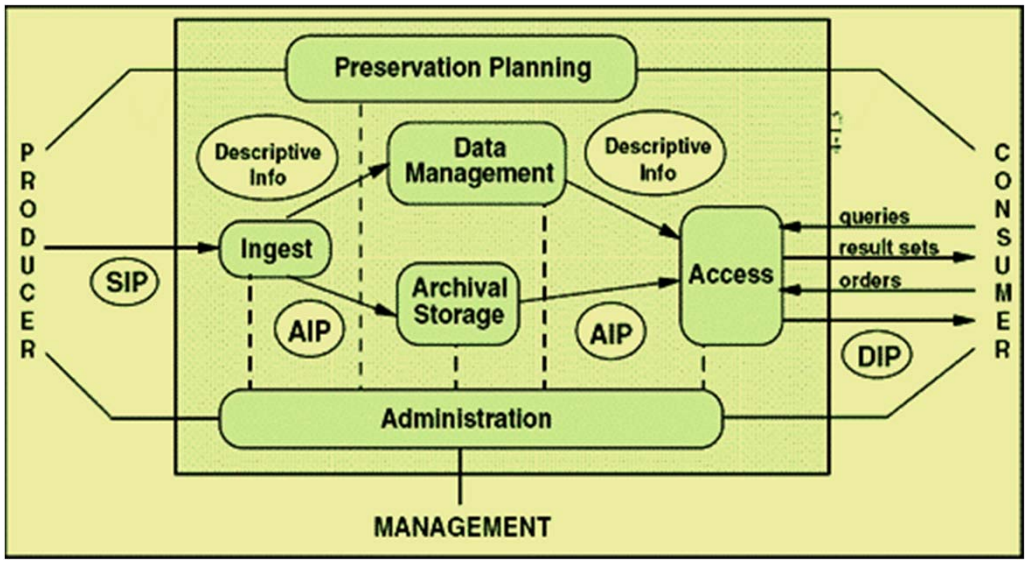
- $10.2 million in State funding over 4 years
- Renew our aging infrastructure and technology systems
- Integrating archive management with collection management
- Innovative interfaces and collaborative spaces
- New website, new Discovery layer



STATE LIBRARY®
NEW SOUTH WALES

OAIS REFERENCE MODEL

STATE LIBRARY®
NEW SOUTH WALES



Hey! How do we make sips?

Migration

Challenge - migrate all the material from ACMS/DAM.

Solution -

- In place storage (all files copied to Rosetta permanent storage)
- CSV export from DAM
- Excel scripts to validate
- CSV to Mets conversion on sandbox
- Ingest METS into Prod

Migration Part 2

STATE LIBRARY®
NEW SOUTH WALES

The DAM Migration Tool

- Meetings and mappings
- PHP (Laravel), MySQL, Web based UI
- Accurate reporting on what has been migrated and what hasn’t for both published and published.
- Uses METS SIPS
- Exceptions are mostly DAM misreporting location

What's in Rosetta so far?

Project	Intellectual Entities	Files
DSM	272	178954
Banks Papers	2248	26231
Rex Hazlewood	121	3237
Macpherson Plates	28	2055
Migration content	316466	2141953
Miscellaneous Ingests	17	303
TOTALS	319152	2352733

The Backlog

Challenges-

- large backlog due to no ingestions while migration happening
- aiming for quick ingest . 5 days from digitisation to presentation (for known formats)
- greater variety of formats - METS/ALTO, oral histories etc
- current BAU - bespoke Python scripts

Content Models

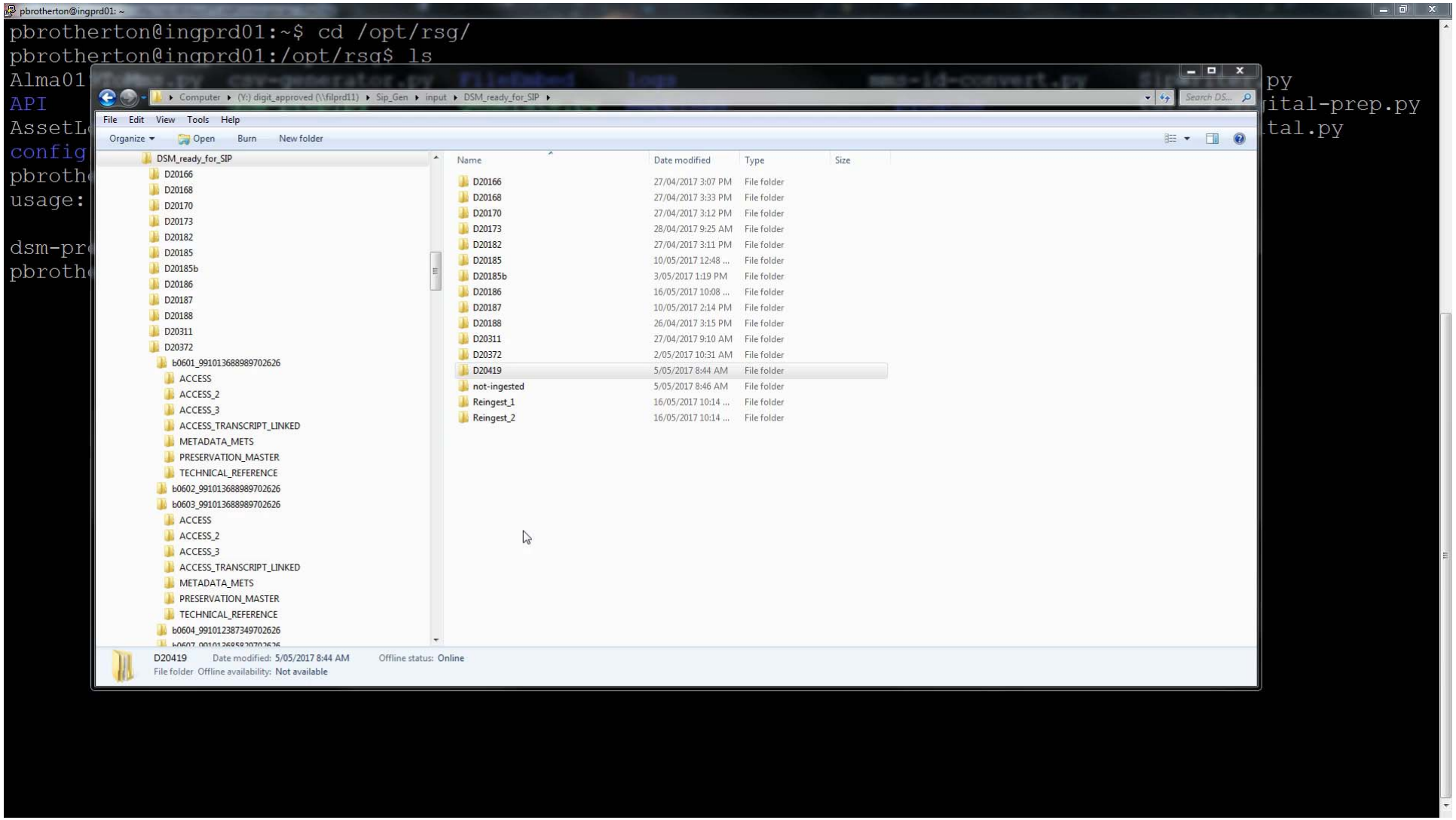
STATE LIBRARY
NEW SOUTH WALES

- Nothing is ingested without a content model
- Sorts of questions that come up:
  - Do we need masters and co masters – how many representations
  - What kind of formats
  - What are the delivery formats
  - Access rights

Solution - What the BAU scripts do ...

STATE LIBRARY®
NEW SOUTH WALES

1. Normalize digitised folders and files
2. Create derivatives (Python scripted Image Magic)
3. Pull down descriptive metadata from either ALMA or Adlib via API lookup
4. Embed descriptive metadata into image files into XMP
5. Generate SIPS
6. Rsync SIPS to Rosetta deposit on network storage
7. Automated ingest jobs set up in Rosetta


pbrotherton@ingprd01: ~
pbrotherton@ingprd01:~$ cd /opt/rsg/
pbrotherton@ingprd01:/opt/rsg$ ls
Alma01
API
AssetL
config
pbroth
usage:
dsm-pr
pbroth
Computer ▸ (Y:) digit_approved (\\filprd11) ▸ Sip_Gen ▸ input ▸ DSM_ready_for_SIP ▸
Search DS...
File Edit View Tools Help
Organize Open Burn New folder
DSM_ready_for_SIP
D20166
D20168
D20170
D20173
D20182
D20185
D20185b
D20186
D20187
D20188
D20311
D20372
b0601_9910136889897026 26
ACCESS
ACCESS_2
ACCESS_3
ACCESS_TRANSCRIPT_LINKED
METADATA_METS
PRESERVATION_MASTER
TECHNICAL_REFERENCE
b0602_9910136889897026 26
b0603_9910136889897026 26
ACCESS
ACCESS_2
ACCESS_3
ACCESS_TRANSCRIPT_LINKED
METADATA_METS
PRESERVATION_MASTER
TECHNICAL_REFERENCE
b0604_9910123873497026 26
Name
Date modified
Type
Size
D20166 27/04/2017 3:07 PM File folder
D20168 27/04/2017 3:33 PM File folder
D20170 27/04/2017 3:12 PM File folder
D20173 28/04/2017 9:25 AM File folder
D20182 27/04/2017 3:11 PM File folder
D20185 10/05/2017 12:48 ... File folder
D20185b 3/05/2017 1:19 PM File folder
D20186 16/05/2017 10:08 ... File folder
D20187 10/05/2017 2:14 PM File folder
D20188 26/04/2017 3:15 PM File folder
D20311 27/04/2017 9:10 AM File folder
D20372 2/05/2017 10:31 AM File folder
D20419 5/05/2017 8:44 AM File folder
not-ingested 5/05/2017 8:46 AM File folder
Reingest_1 16/05/2017 10:14 ... File folder
Reingest_2 16/05/2017 10:14 ... File folder
D20419 Date modified: 5/05/2017 8:44 AM Offline status: Online
File folder Offline availability: Not available

0:00

Here is an example of a digitisation of books from our David Scott Mitchell (DSM) collection. They are divided into work orders, this particular work order D20419 contains 51 books.

Opening up one of the books you can see that we have a number of subfolders containing files of different types. Each folder will end up as a separate Representation within the IE in Rosetta.

0:40

Now we are going prepare the files and folders so they are in a structure which our general purpose SIP generator script requires.

1:05

Here you can see the results. The Representation folders have been renamed, we have generated meta.json placeholders which contain the catalogue ID. In this case an MMS ID from Alma.

1:33

We have file meta.json files in each Representation folder. We also have files to indicate our Preservation Type and Representation Code.

1:50

Now we have the source files ready we can generate the SIPs. Here we will be

Retrieving descriptive metadata from our Library Catalogue via web services (Alma).

Embedding the metadata into the tiff and jpeg files as XMP.

Generating the SIPs.

2:30

Here you can see the log being produced by the script. We generate logs in a structured json format so they can be easily processed later on for reports and analytics.

3:11

Here is one we prepared earlier. You can see the files are now zipped up and we have a CSV structured as Rosetta requires for SIP deposits. You can see the metadata which we pulled down from the catalogue and other required data such as filenames and access rights.

Limitations

- If something goes wrong we need to reprocess the entire batch again
- Manual intervention is required by a developer to configure and start the scripts
- Manual intervention is required to rsync the SIPs to Rosetta deposit storage
- Manual monitoring of the Rosetta ingestion is needed to track which SIPs fail
- Custom scripts need to be written to get source files and folders into a structure which can be processed

The future …. PANDA

- **P**reservation **AN**d **D**igital **A**ccess
- Full automation for business users
- Web GUI
- Born digital pre-conditioning
- Big focus on reporting and analytics

STATE LIBRARY®
NEW SOUTH WALES