

REPOSITORY MANAGEMENT

Stuart Hungerford

Background

- Based at ANU Supercomputer Facility
- Host of APAC National Facility and MDSS
- Supporting projects with large or complex dataset needs
- Now supporting APAC Data Intensive Projects

Projects Supported

- PARADISEC
- ACLA
- Reciprocals
- EPINET
- Others

Repository Management

- Look at some issues in managing DIP project repositories
- Use PARADISEC as example project and repository
- Refer to Linda Barwick's presentation for background

Repository Success Factors – Community

- Successful project repositories often have strong communities
- Well defined standards for file formats
- Well defined standards for metadata
- Defined best practices for curation
- Defined models for rights management and access

PARADISEC – Community

- Active endangered languages and cultures community
- Standard for digital audio preservation (BWF)
- Standard for XML metadata harvesting
- Clear and rigorous workflow including curation
- Strong rights management needed

Repository Success Factors – Expertise

- Within communities much expertise available
- Successful project repositories tap this expertise
- Curation, preservation formats, DOI's, workflow, metadata
- Within Australia: Archives, Library, ScreenSound, Universities, other

PARADISEC – Expertise

- At project planning stage took advice
- NLA, ScreenSound, National Library, stakeholders

Repository Technology Enablers

- Unified storage view of preservation copies
- Fast networking for data and metadata access
- Standards compliant web access

PARADISEC – Technology Enablers (I)

- National Facility MDSS
- 1.2PB (soon 6PB) nearline storage
- Refer to Ben’s presentation for details
- GrangeNet research network
- 10Gb partner connections
- Subscription not volume model
- MDSS system topologically close to ACT routers
- Network bottlenecks now inside institutions

PARADISEC – Technology Enablers (II)

- Standards compliant web access to data and metadata
- Universal GUI for all stakeholders
- Prefer agile toolkits and approaches
- Prefer REST approach to web access
- Foundation for web services access
- Foundation for value-added projects with datasets

Repository Automation

- Successful repositories rely on automated operations
- Manual workflow too error prone
- Must have support for oversight and reporting
- Must be able to answer questions like
 - “what is the current situation?”
 - “how can we improve the workflows?”
 - What if questions

PARADISEC Automation (I)

- Nightly copy of new or changed files from ingest site(s)
- Weekly generation of repository reports: contents, metrics, audit
- Logging of repository events: new files, changed files, removed files
- Reconcile with off-site metadata stores

PARADISEC Automation (II)

- Automatically check repository data against audit “rules”
- For example: properly formed file name
- For example: BWF file has MP3 quick access version
- Currently approximately 40 audit rules
- Audit failures feedback into improving ingestion workflow

Repository As Project Foundation

- Well managed repository can be basis for value-ad projects
- One repository—many interfaces
- Foundation for e-research type projects

PARADISEC Project Foundation

- Basis for individual researchers work
- Now basis for ARC e-research project in annotation

PARADISEC Repository Futures

- More ingest sites
- Ingest born-digital materials
- Ingest video
- Streaming support