



**ERPANET @SAA2004:**  
**Trusted Repositories**  
*SA Pre-Conference Workshop*  
*30 August 2004, Glasgow*

Seamus Ross  
ERPANET



Presentation based on  
ERPANET's Rome Workshop  
See Rome presentations at:

<http://www.erpanet.org>



# Digital Preservation Requires

- What are the features of a ‘trusted digital repository’?
- What current implementations may be rightfully called “trusted digital repositories”?
- How are the concepts of reliability, authenticity and trustworthiness interpreted in different contexts and why?
- How should the roles and responsibilities of the many stakeholders be addressed?
- Is the potential of trusted digital repositories currently being adequately exploited?
- What issues have not yet been addressed in trusted digital repositories implementations and research?
- Is it possible and/or necessary to agree on one definition of trusted digital repository?
- How do different communities see trusted digital repositories?
- Creating and maintaining the trust of their user communities overtime and in the face of changing technologies.

# Preservation Models: OAIS

- OAIS = Open Archival Information Systems
- Key players in development
  - National Space Science Data Centre
  - Consultative Committee for Space Data Systems
- Now an ISO Standard
- Premises Underlying OAIS
  - Data are irreplaceable (esp observation)
  - Data and associated metadata must be moved across technologies
  - Representations and formats will change
  - Lack of consensus on adequate metadata standards

# Key OAIS Objectives

- Objective
  - recognised no framework for developing digital archive standards
  - need for a reference model
  - recognise the hybrid nature of archives
  - collaborate with archival community
  - focus on data resulting from space missions
  - near-term and indefinite storage of digital data
  - independent of implementation model
  - address full range of archival processes

# Process of Development

- Examine other models
- Define Data Archiving
- Define functional areas (FAs) including ingest, storage, access, and preservation
- Define interfaces between FAs
- Define a set of data classes
- Formal representation methods

# OAIS Overview

- Manages ingest of Information Packages from creators
- Defines the communities needing the Information
- Reflects needs of identified user community
- Enables preservation in an understandable way
- Uses documented policies and procedures

## Advantages of OAIS

- Provides a model where one was lacking
- Facilitates procurement of systems
- Enables interoperability between OAIS compliant systems
- Supports the migration task
- Lays out a minimum set of responsibilities



## What does it support?

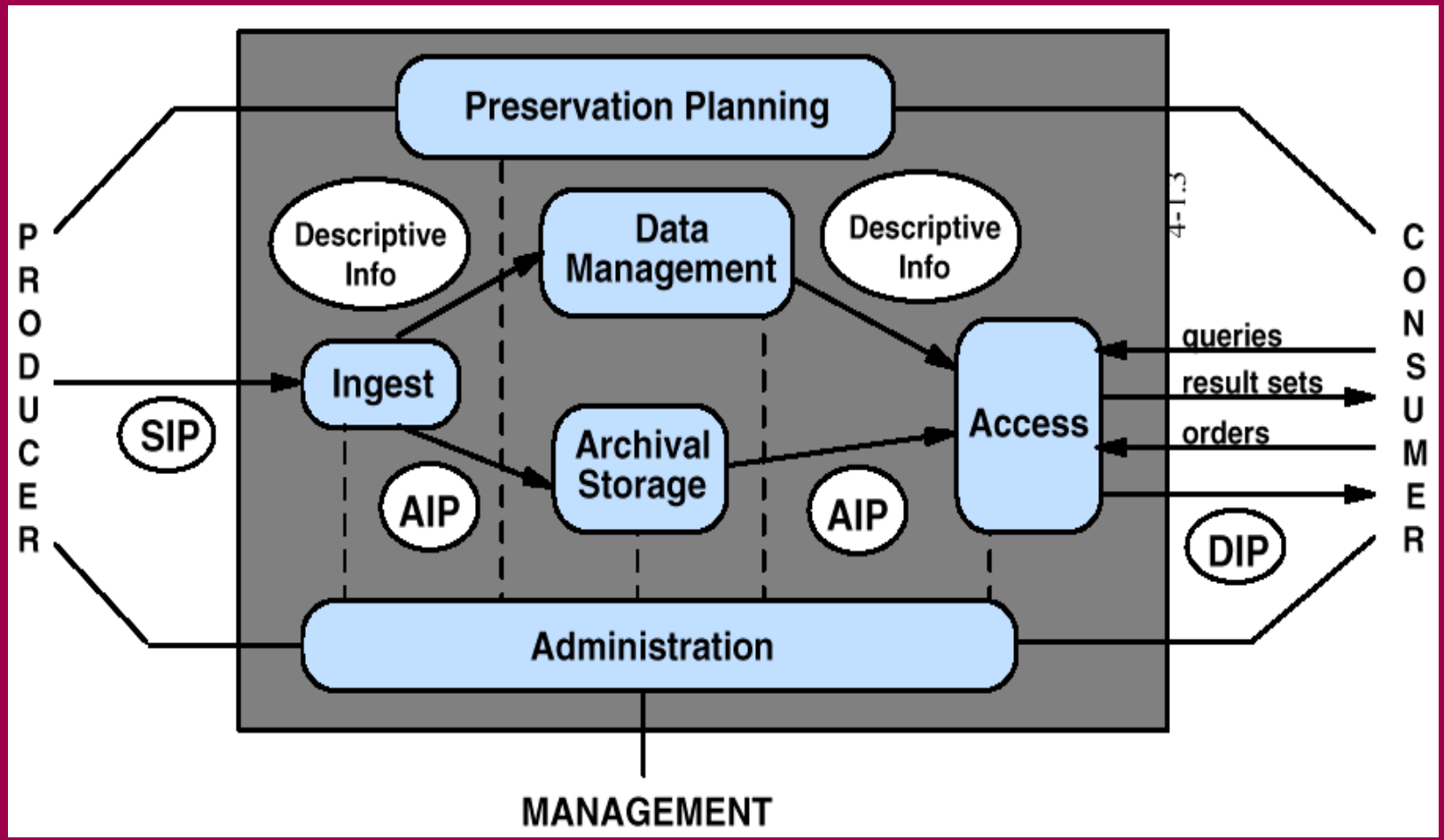
- Resource Information about objects
  - Submission Information Package
  - Archival Information Package
  - Dissemination Information Package
- Relates activities of producer, manager and consumer
- Supports Functional Entities

# OAIS Functional Entities

- Ingest
- Archival Storage
- Data Management
- Administration
- Preservation Planning
- Access



# OAIS Functional Entities



## Who is working with OAIS

- Archive & Library Community
  - Koninklijke Bibliotheek (KB) through NEDLIB—design and architecture of Deposit System for Electronic publications
  - CEDARS
  - NARA and the San Diego SuperComputer Center
  - National Space Science Data Center
  - Pharmaceutical & Aerospace Industries
  - French Space Agency for its plasma physics archive

# What must a Repository Do

- Handle a wide array of digital media types
- Be Secure
- Guarantee authenticity of the objects it holds
- Protect Integrity (from intended and unintended harm)
- Enable verification
- Ensure stuff ingested into the archive can be output (e.g. be accessible)
- Self-contained (in operation)
  - Must not rely on external infrastructure or services
  - Maintain all documentation in-house
  - Have disaster recovery functionality built-in

# Repositories must be trusted

- Processes:
  - Workflows
  - Operation (management of integrity, authenticity, intelligibility, and accessibility)
  - Automation (e.g. ingest, management, publication)
  - Documentation of procedures
  - Auditability
- Architecture and Implementation

# Flexible Technical Infrastructure

- Hardware and software elements must be
  - Sustainable
    - Financially
    - Reasonable levels of complexity
    - Management
  - Open
    - Modular (flexibility to dump obsolete HW and SW)
    - Migratable (possibility to change to HW and SW)
  - Secure (no unauthorized access)
  - Reliable (no data loss)
  - Available (the data is at hand when you need it)

# Authenticity

- Requires control of ingest and its verification
- Depends on immutability of the data store
- Migration may destroy original byte stream
  - archives and stakeholders must identify significant properties and validate their migration
- Support Audit of the chain of custody, process history, and the descriptions of the migration processes
- Provide mechanisms to enable use (including access to context information and rendering)



# Repository Operation

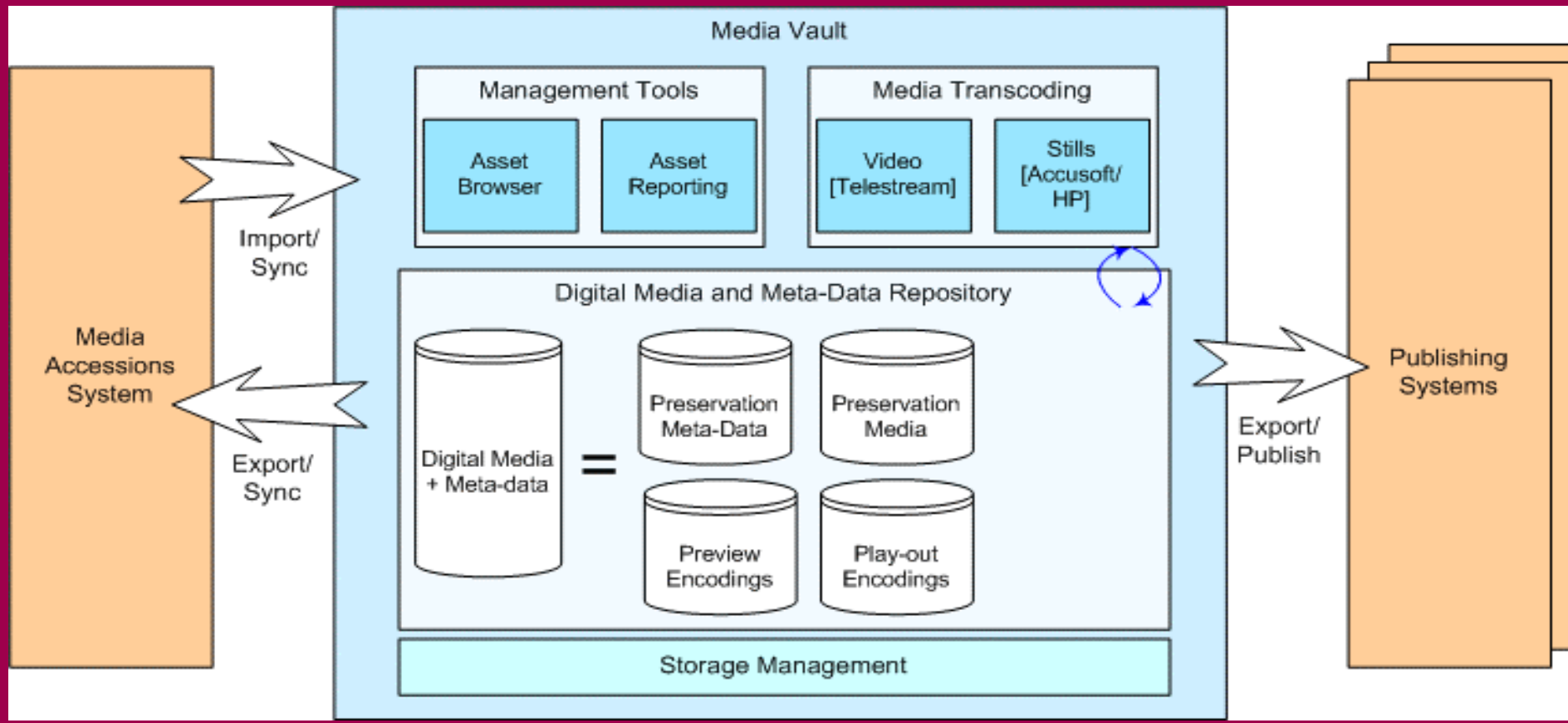
- Change will be a feature of repositories
  - Storage technologies
  - Services, close down of some and initiation of others
  - Workflows
  - Verification mechanisms
  - Migration, refreshing, emulation

# Automation

- Huge quantities of materials to ingest and manage
- Automation of workflows allow integration of independent services
- Standardized logging/record creation
- Reduce human intervention
  - Cheaper
  - Less error prone
  - Enables higher level of security and reliability
- Intensive test and verification needed
  - Mistakes are very costly (financially but more importantly in terms of trust)



# Media Vault (ARKive)



# National Archives of Australia

- NAA follows OAIS Framework
- Uses an open source-based solution
- Focuses on use of xml
- Uses open and well-documented data formats
- Not creating a digital archive, but a preservation approach
- Need to address: appraisal/selection, transfer, description, and retrieval and access

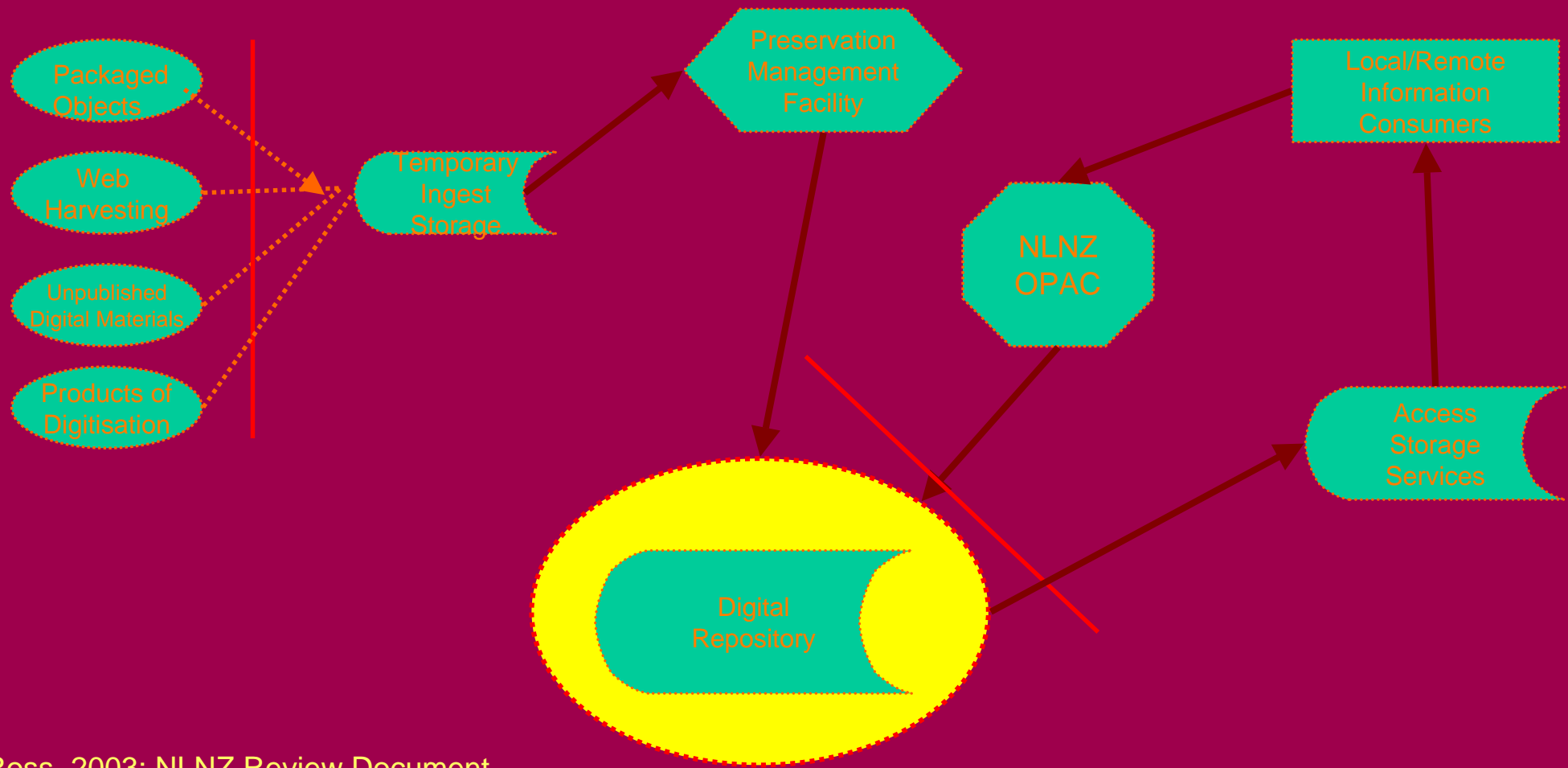


# Preservation System

- 3 separate components
  1. Quarantine
  2. Preservation
  3. Storage
- All components physically separated from each other and all other NAA networks
- Access to hardware restricted to digital preservation staff



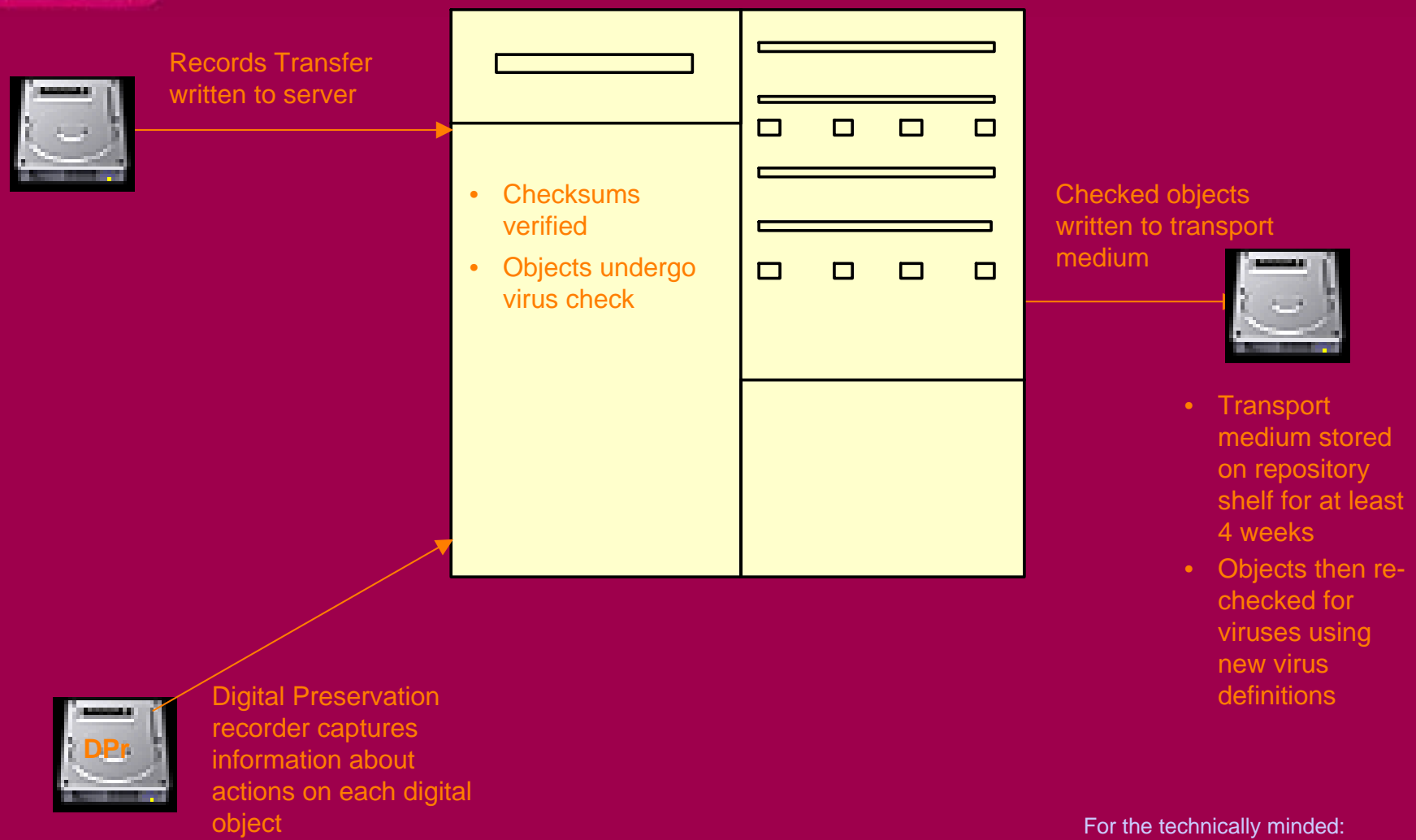
# National Library of New Zealand: Proposed Digital Repository Structure



Ross, 2003: NLNZ Review Document.



# Quarantine server



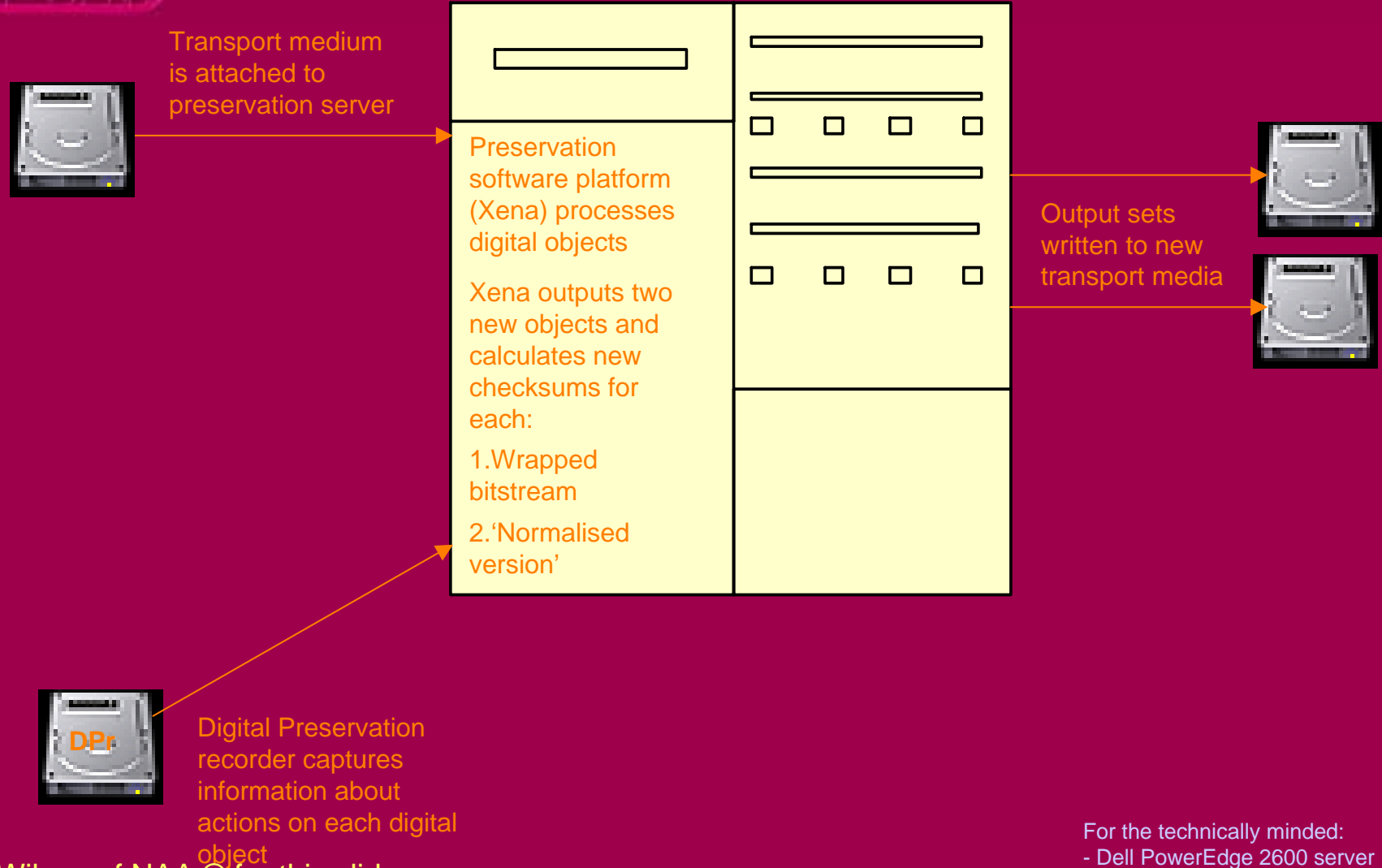
Thanks to Andrew Wilson of NAA © for this slide.

For the technically minded:

- Dell PowerEdge 2600 server
- 2 x 2GHz processors
- .7Tb disk store
- independent UPS



# Preservation server



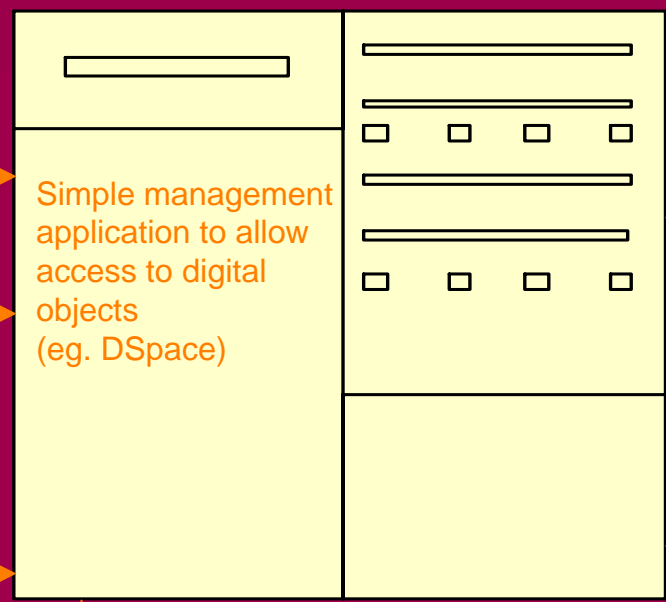
Thanks to Andrew Wilson of NAA © for this slide.

ERPANET & HATII -- Seamus Ross: Trusted Digital Repositories

- For the technically minded:
- Dell PowerEdge 2600 server
  - 2 x 2GHz processors
  - .7Tb disk store
  - independent UPS



# Digital Repository

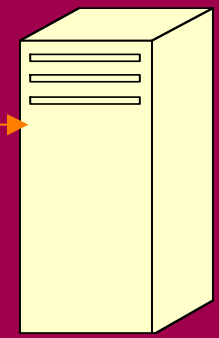


Transport media are attached to repository server



Simple management application to allow access to digital objects (eg. DSpace)

Third copy on digital tape which is stored offsite



Digital Preservation recorder captures information about actions on each digital object

Copies written to new media for access



To Access



2 copies on RAID storage  
- Configured as RAID 10  
- Automated, regular, frequent verification of checksums

- For the technically minded:
- Dell PowerEdge 2600 server
  - 2 x 2GHz processors
  - .7Tb disk store
  - fibre channel between server and RAID
  - independent UPS

Thanks to Andrew Wilson of NAA © for this slide.

# ARELDA

## Archiving of Electronic Digital Data and Records

- Goal: Finding long-term solutions for the permanent archiving of digital records in the Swiss Federal Archives
- Indispensable for the long-term execution of the Federal Archives Act
- Development costs 2001 – 2008: ~ 11 Mio €
- Operational costs from 2005: 2.5 – 3.3 Mio € per year (expected growth: 20 TB/yr net)
- Today's project team: 7 people (4 CS engineers)

Thanks to Stefan Heuscher of Swiss Federal Archives © for this slide.



# NARA's Strategic Response – ERA Requirements

- **Persistent**

- To manage and access the records over time.

- **Authentic**

- To ensure that these are the original records
- Records that are created with attached documentary information

- **Scalable**

- To grow and adapt to increasing volumes and evolving types of electronic records
- To serve a variety of user groups (e.g. rich service layer)

# Technical Challenges of ERA

- Receives, stores, preserves, and provides access to electronic records, regardless of type, format or media.
- Receives, preserves, and store electronic records in a manner and environment appropriate to their sensitivity level.
- Stores electronic records in a manner that allows for maximum possible independence from specific hardware and software infrastructures.
- Supports high availability.
- Provides viable long-term storage for electronic records.

## NARA argues: repositories must

- Find records based on searches of descriptions of records
- Search the electronic records themselves.
- Accurately reproduce and output electronic records.
- Provide certified copies of electronic records.
- Manage requests for review of restricted materials.
- Implement the results of electronic records reviews.
- Enable users to request and receive assistance from archivists or librarians.

## Are there tools

- LOCKSS (Lots of Copies Keep Stuff Safe)  
<http://lockss.stanford.edu>
- Fedora --**FEDORA (Flexible and Extensible Digital Object and Repository Architecture)**  
<http://www.fedora.info/>
- DSPACE--<http://dspace.org/index.html>
- Digital Asset Management Systems ???

# Digital Asset Management Systems (1)

- What new opportunities will a DAM system enable the institution to create? How will the institution measure whether or not DAM has achieved the objectives?
- What functions of DAM systems are particularly well suited to the needs of the institution?
- How can the institution ensure staff buy-in to DAM technology?
- What will be the cost-benefit ratio?
- Which DAMS technologies best fit the institutions requirements? How will the selection process be documented?
- Has the institution established that the target DAM system can be optimised for the data types which the organisation handles, that it supports adequate user profiling and that the metadata categories supported are adequate?
- What impact will the introduction of a DAM solution have on organisational thinking about and use of digital content?

## DAMS/CMS (2)

- What might the implications be in collaborating with other institutions to share a DAMS?
- What obstacles might be encountered when attempting to introduce DAM technology? How can these be overcome?
- What metadata are required to support the institution's application of DAM technology? How will the metadata be acquired and implemented?
- DAMS are based on a combination of technologies and methods, including software applications and policies and procedures. Have those elements that are software-based, and those concerned with policies and procedures been identified?
- Have plans to develop, test, disseminate, and validate the application of these policies and procedures been established?
- Will a DAM system allow recognition of the economic, educational, or intellectual value of digital assets that have hitherto been overlooked?

## DAMS/CMS(3)

- Will a DAM system allow the institution to exploit the economic value of its digital content?
- What risks to the institution's digital content are posed by the use of DAM technology?
- How will DAM technology be integrated with existing systems like digitisation systems?
- As DAMS cannot help protect intellectual property rights, what are the IPR implications of establishing a DAM system for the institution?
- As most DAMS do not necessarily provide long-term preservation of digital assets. How will the organisation address this problem?

# Are there Repository COTS?

- Yes & No
  - No single out of the box solution to all aspects
  - No solution that you can adopt and rollout
  - There are increasing numbers of models covering aspects of the problem

# Challenges

- Is it possible and/or necessary to agree on one definition of trusted digital repository?
- How do different communities see trusted digital repositories?
- What current implementations may be rightfully called “trusted digital repositories”?
- How are the concepts of reliability, authenticity and trustworthiness interpreted in different contexts and why?
- How should the roles and responsibilities of the many stakeholders be addressed?
- Is the potential of trusted digital repositories currently being adequately exploited?
- Will Archives become conservators of the original’s creator authenticity?
- What aspects are essential and what are incidental in determining a records value? (e.g. what loss is acceptable)