

# Digital Preservation Research Initiatives at NLNZ



SAA Research Forum  
Austin, Texas 11 August 2009

Steve Knight, Programme Director Preservation Research and Consultancy  
National Library of New Zealand



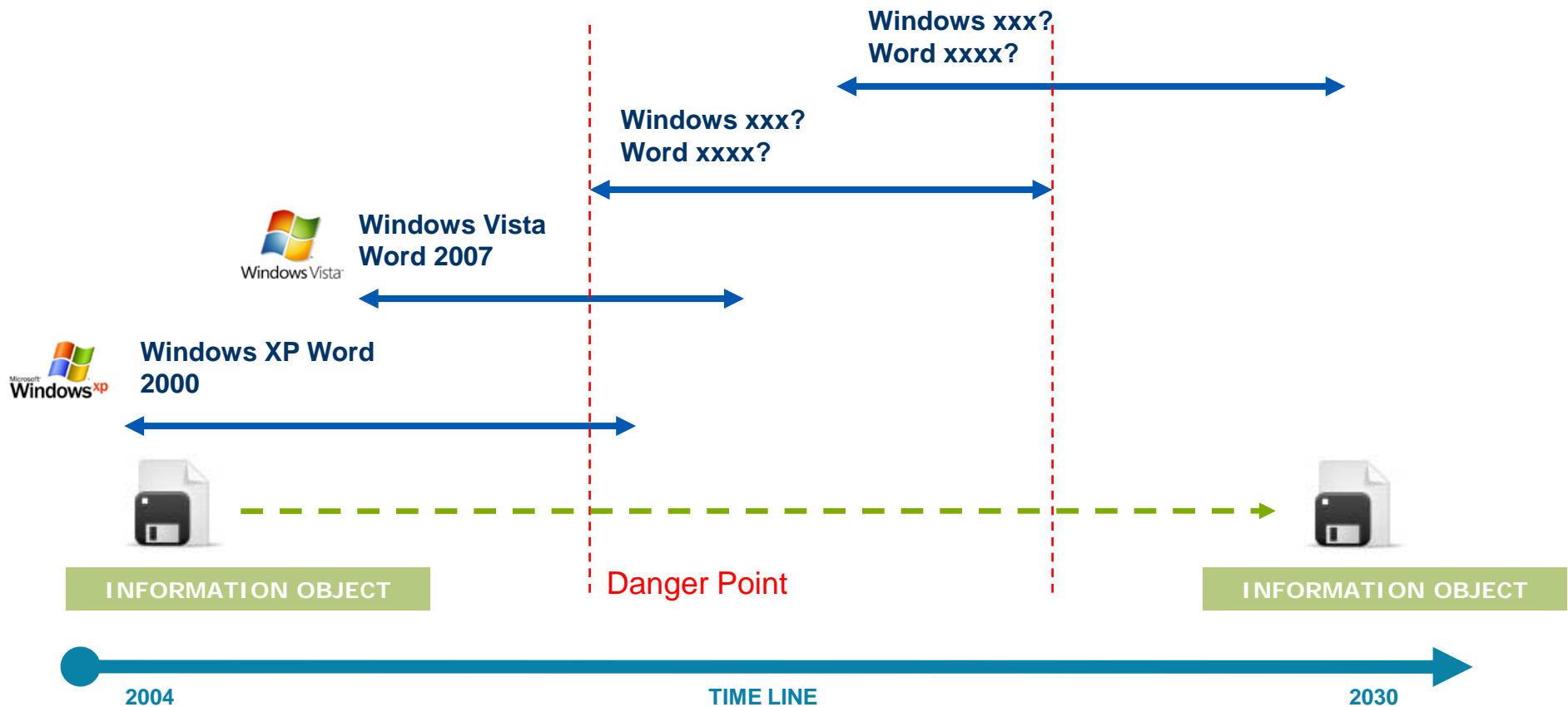
# Today

- A bit about the National Digital Heritage archive project at NLNZ
- A look at how we're starting to model preservation risk management
- Some comments on the state of digital preservation



# Technology as the enemy

## The evolution of technology environments





# What is the real digital preservation problem?

Garrett, J. & Waters,  
D. (eds). (1996)

Preserving digital  
information ...

*‘ the problem of preserving digital information for the future is not only, or even primarily, a problem of fine tuning a narrow set of technical variables. It is not a clearly defined problem ... rather, it is a grander problem of organizing ourselves over time and as a society to maneuver effectively in a digital landscape. It is a problem of building ... the various systematic supports ... that will enable us to tame the anxieties and move our cultural records naturally and confidently into the future.’*



# What do we do this for?

Why should national  
libraries or everyone  
here care?

*“A National Library is a place where a nation  
nourishes its memory and exerts its imagination  
where it connects with its past and invents its  
future.”*

Pierre Ryckmans. 1996. “Perplexities of an  
electronically illiterate old man.”

Quad-rant, September 1996, No 329.



# What do they do this for?

## Public Records Act 2005



A government  
archiving point of  
view

*“through the systematic creation and preservation of public archives and local authority archives, to enhance the accessibility of records that are relevant to the historical and cultural heritage of New Zealand and to New Zealanders' sense of their national identity .”*



# What we think we're doing

Some milestones

There is still a long  
way to go

**Jul03** : Preservation Metadata Schema and Logical Data Model (iteration2)

**Jul04** : NDHA Programme established

**Sep04** : Metadata Extraction Tool

**Sep05** : Object Management System (OMS)

**Sep05** : Interim Electronic Legal Deposit (IELD) Online Submission mechanism/processes

**Nov05** : Sun Centre of Excellence for Digital Futures in Libraries announced

**May04** : NDHA Business Requirements Specifications

**Nov05** : NDHA Functional Requirement Specifications

**Sep06** : Web Harvesting Web Content Tool (WCT)

**Mar07** : NDHA / DigiTool Gap Analysis completed

**Oct08** : Phase 1 Rosetta delivered

**Feb09** : NDHA launched at NLNZ

**Nov09** : Phase 2 Rosetta delivered

**MAR10** : INITIAL DREAM OF NDHA COMPLETED



# What we think we're doing

## Components of digital preservation

Storage  
psychology  
Processes/strategies

This is only the  
beginning

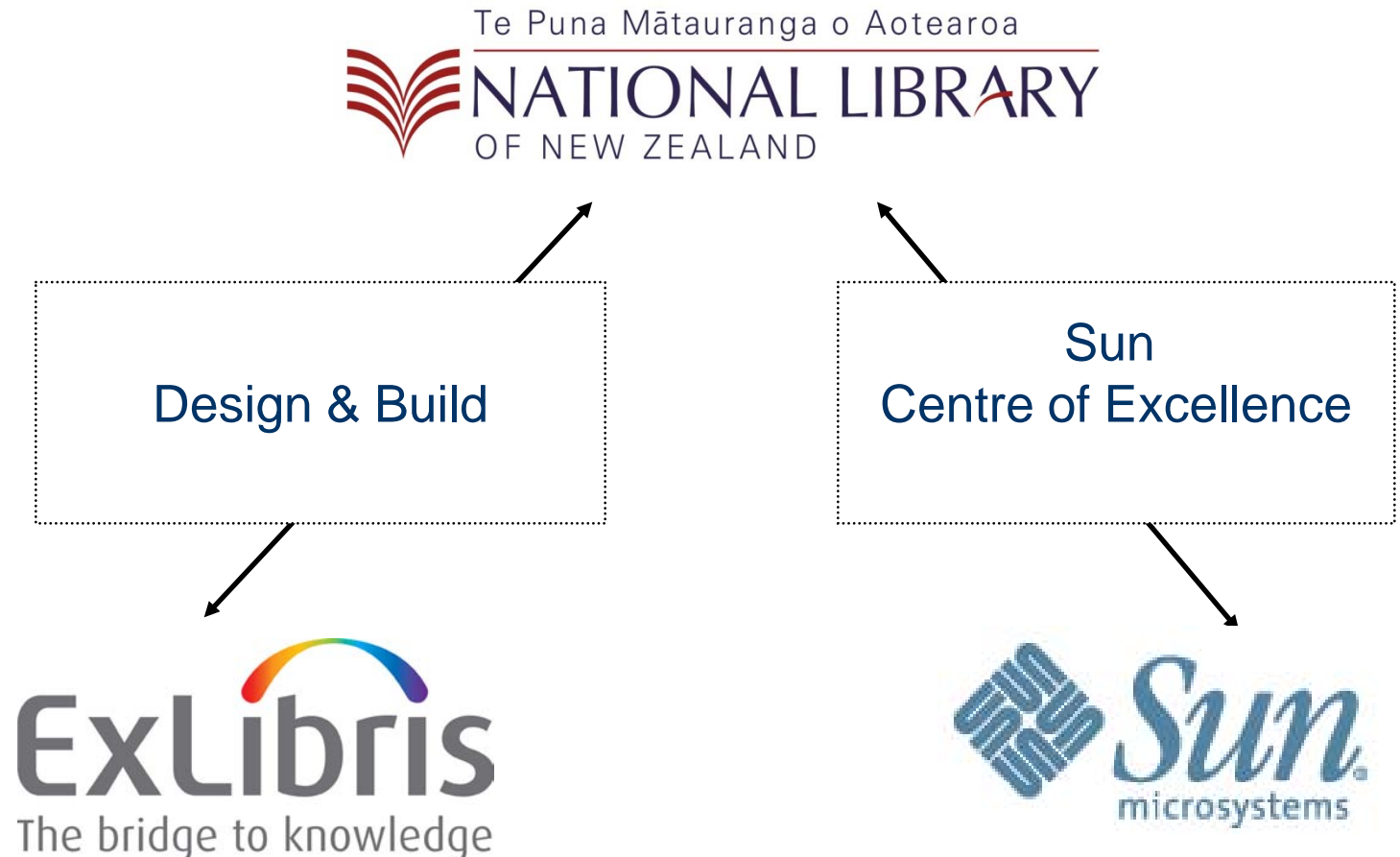




# Collaboration

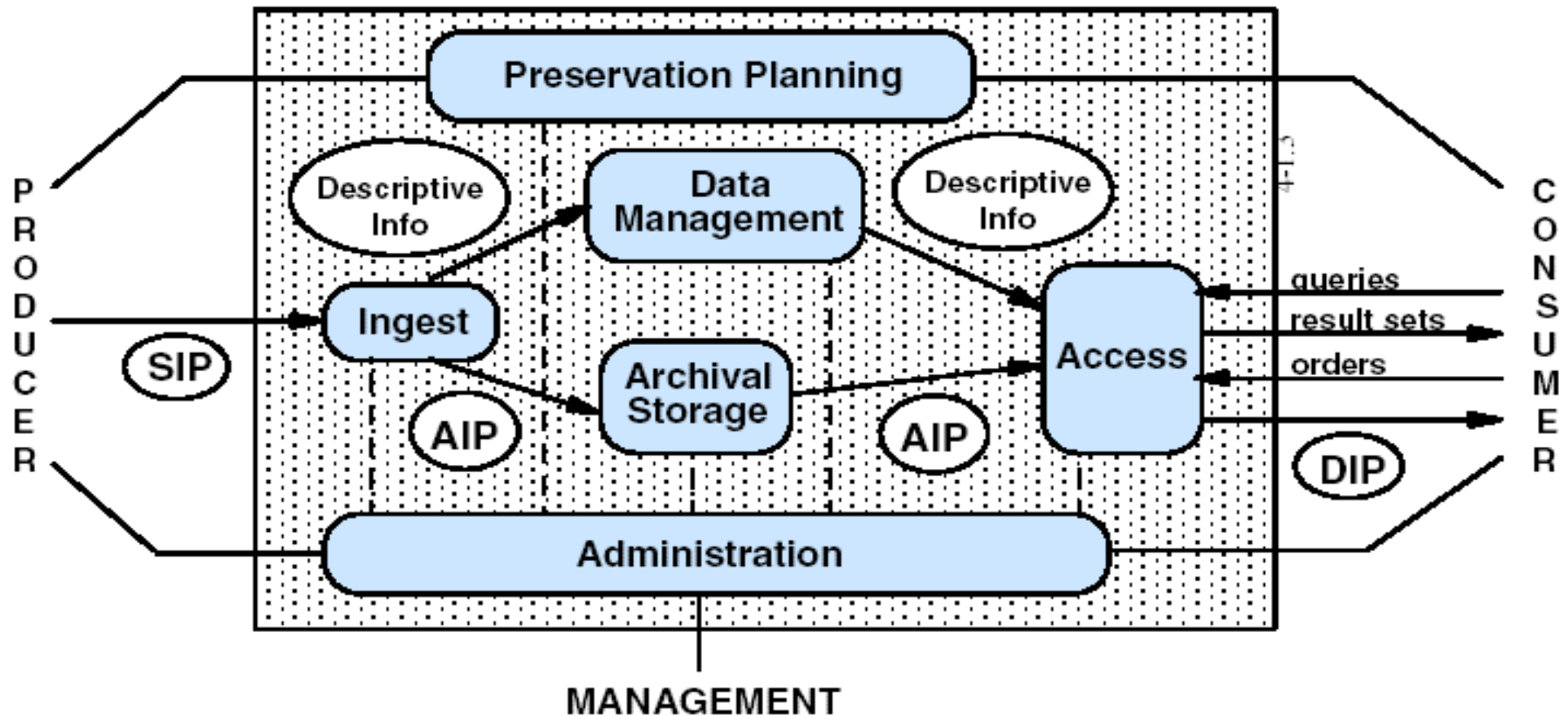
## Partnership

The NDHA Programme will be successful and delivered in a timely and cost effective manner



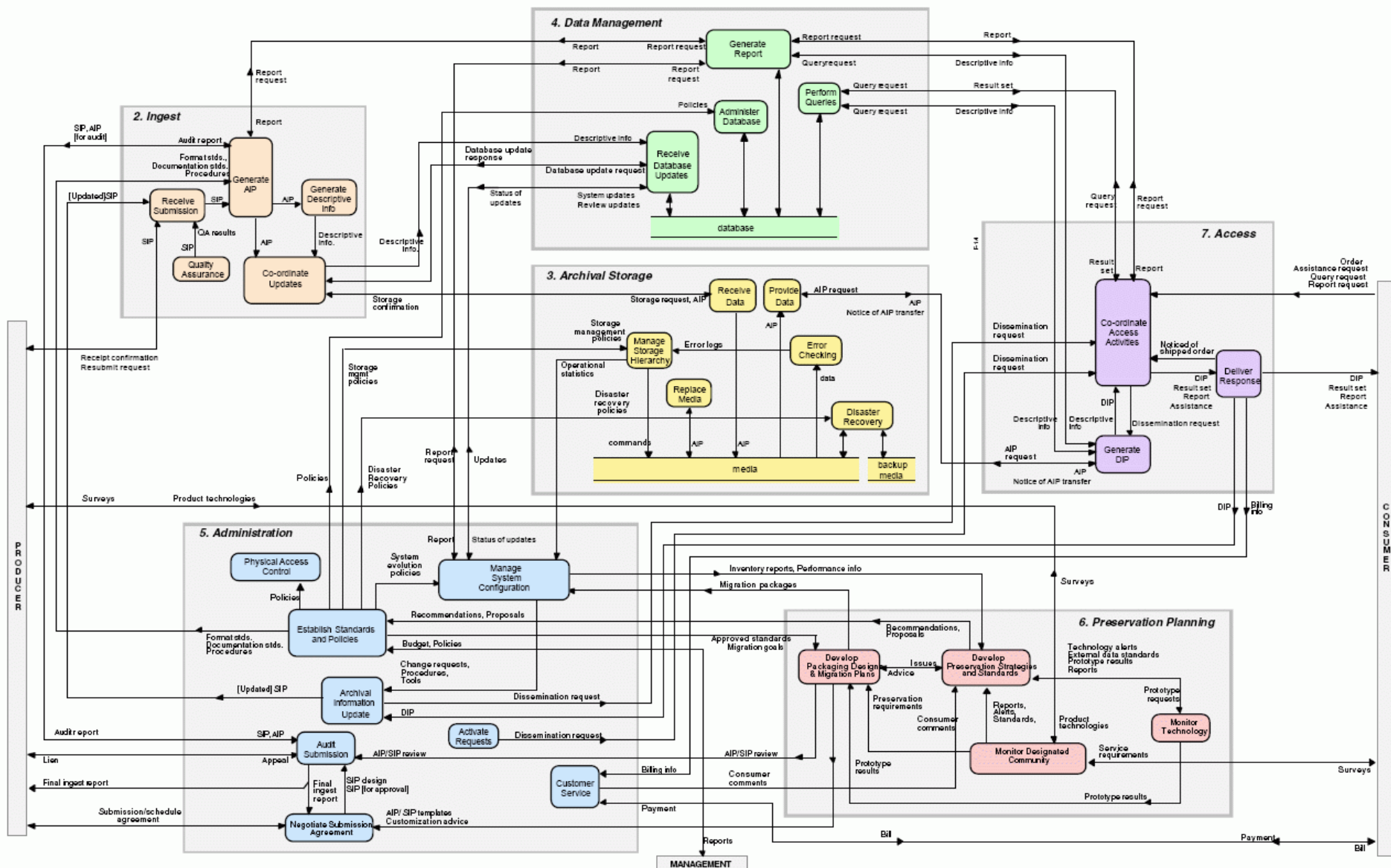


# OAIS Reference Model





# OAIS base map





# Rosetta Functionality Day 1

## Phase 1 Specs

From producer management → workflow automation → delivery, audit trails & reporting

- User management
- Producer management
- Deposit 1
- Deposit 2
- Validation stack
- Intellectual Entity (IE) data model
- Submission Information Package (SIP) submission
- SIP processing
- Deposit registration
- Technical analyst
- Workbench
- Consolidated appraisal workbench
- Rosetta transformers
- Deposit Application Programme Interface (API)
- Audit & provenance
- Process management
- User management API
- Permanent repository
- Delivery
- Meditor
- Reports
- Back office configuration



# Phase 2 Delivery

Phase 2  
Specs

**Enhanced risk  
management  
functionality**

**Mmmmm**

- Format Library
- Risk analysis
- Preservation action
- Enhanced set import/export management for preservation actions
- Maintenance and management functions in Staging NOT permanent
- Enhanced configurability

Preservation planning and action are in Phase 2 to allow for extended requirements analysis prior to development



# Integration

## Integration work stream

It's not all about the  
Digital Preservation  
System

- Deposit applications development
- Existing collection management systems integration
- Browser based content delivery systems development
- Existing resource discovery and delivery systems integration
- Reporting systems
- Common services integration
- Data migration



**INDIGO**

Forms ...

Romanic: indicum,  
Indicus

Spanish: indico

Portuguese: endego

Dutch: indigo

NDHA: in dey go

## Internal Submission Application

- Submission Information Package (SIP) Creation Tool (Templates, Hotkey support)

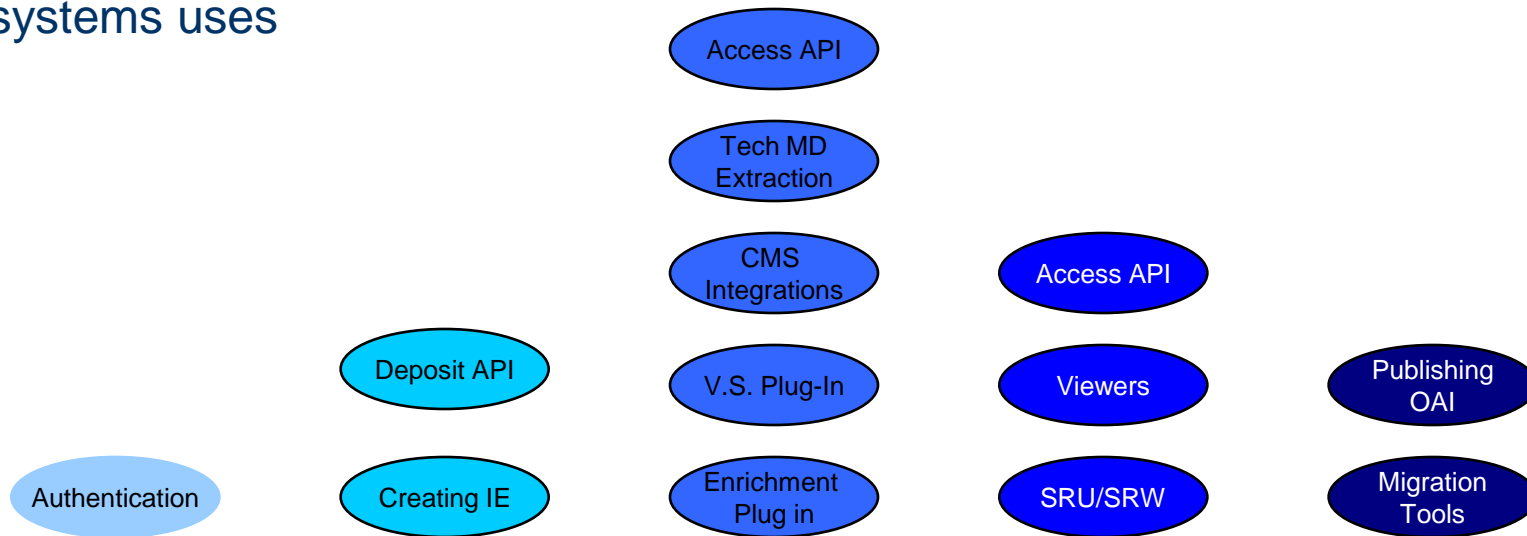
## Packages up

- Files (supports complex digital objects)
- Metadata (Structure map creation – METS)
- Digital object structure – multiple representations
- Fixity generation (MD5)
- Links to descriptive record – CMS integration
- Links producer records
- Submits SIP to the NDHA



# Integration Points

## Customer systems uses



## Digital Preservation



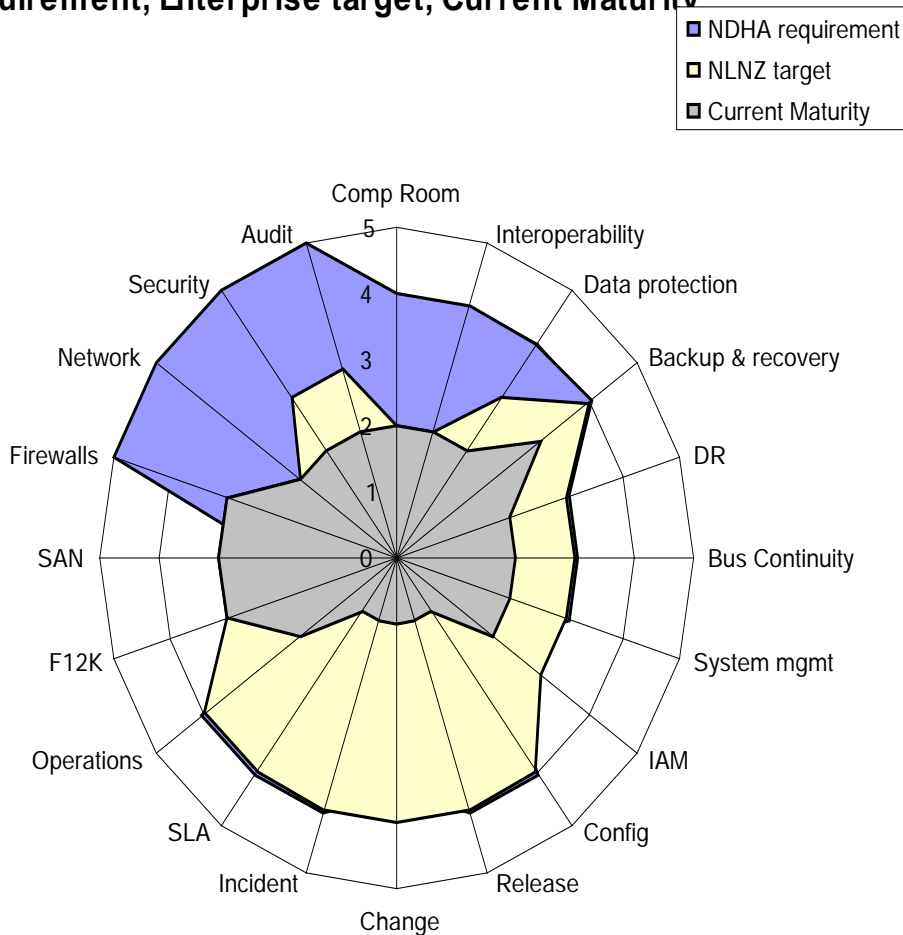


# Technology Infrastructure

## Maturity

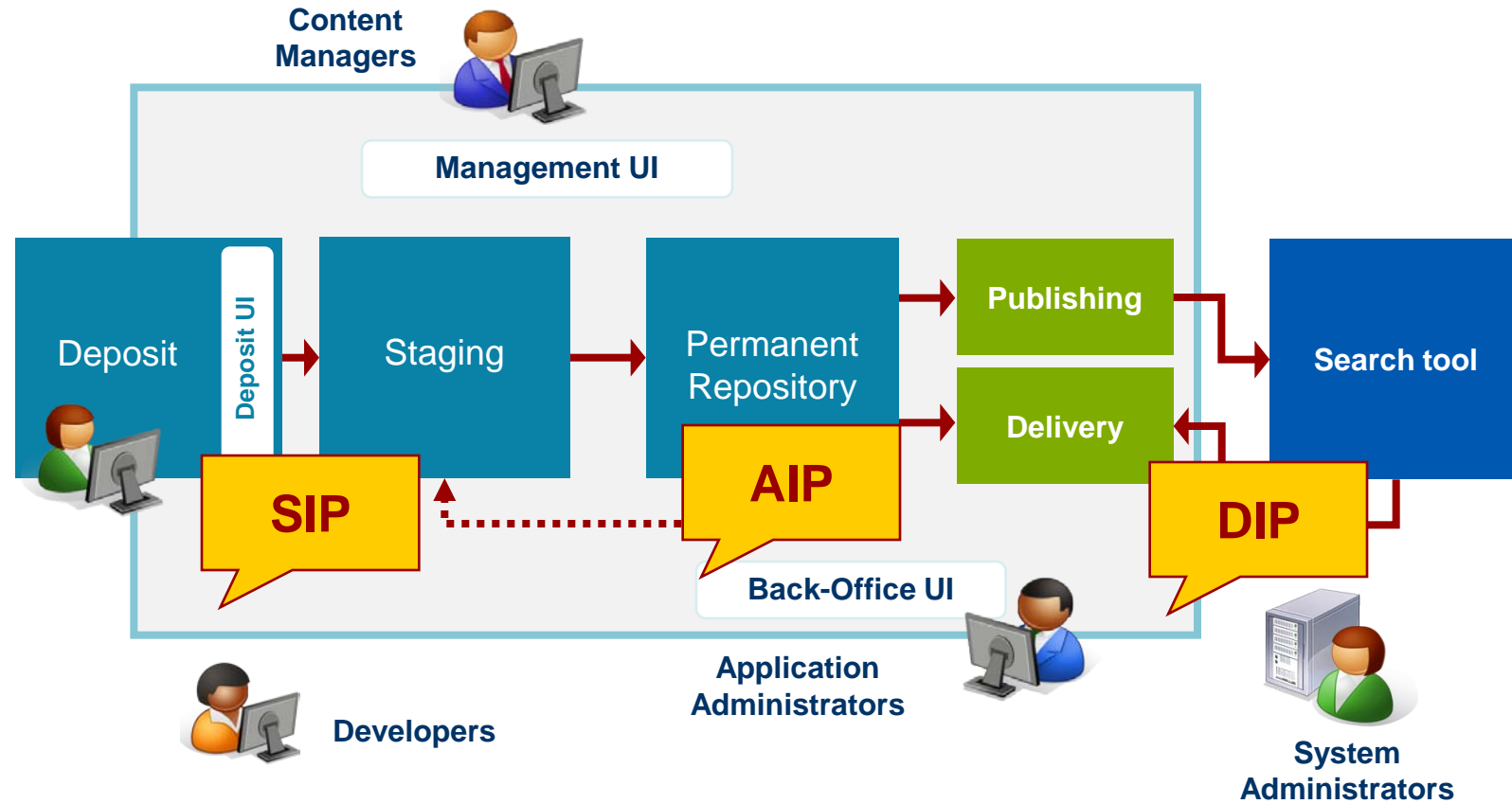
How ready is our infrastructure for digital preservation?

### NDHA requirement, Enterprise target, Current Maturity



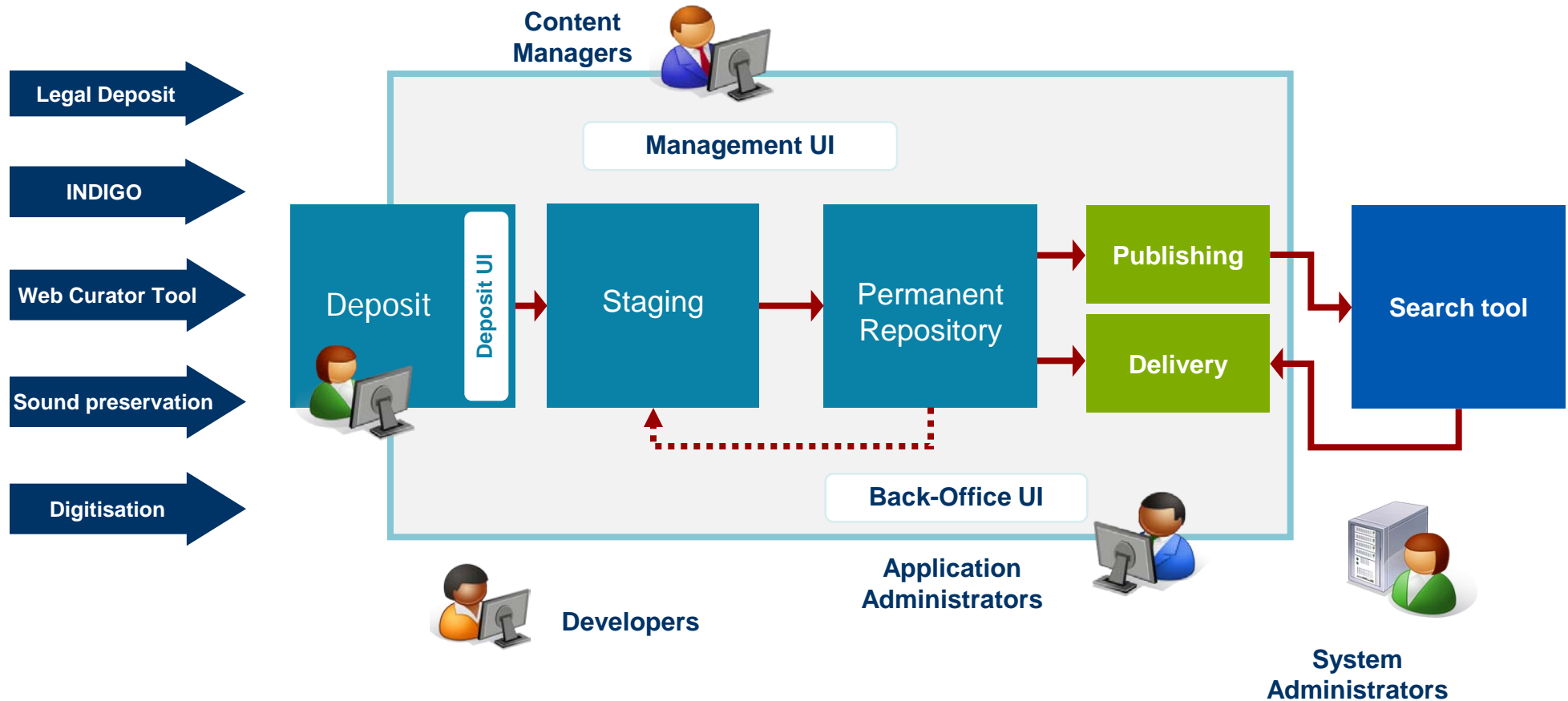


# SIPs, AIPs & DIPs



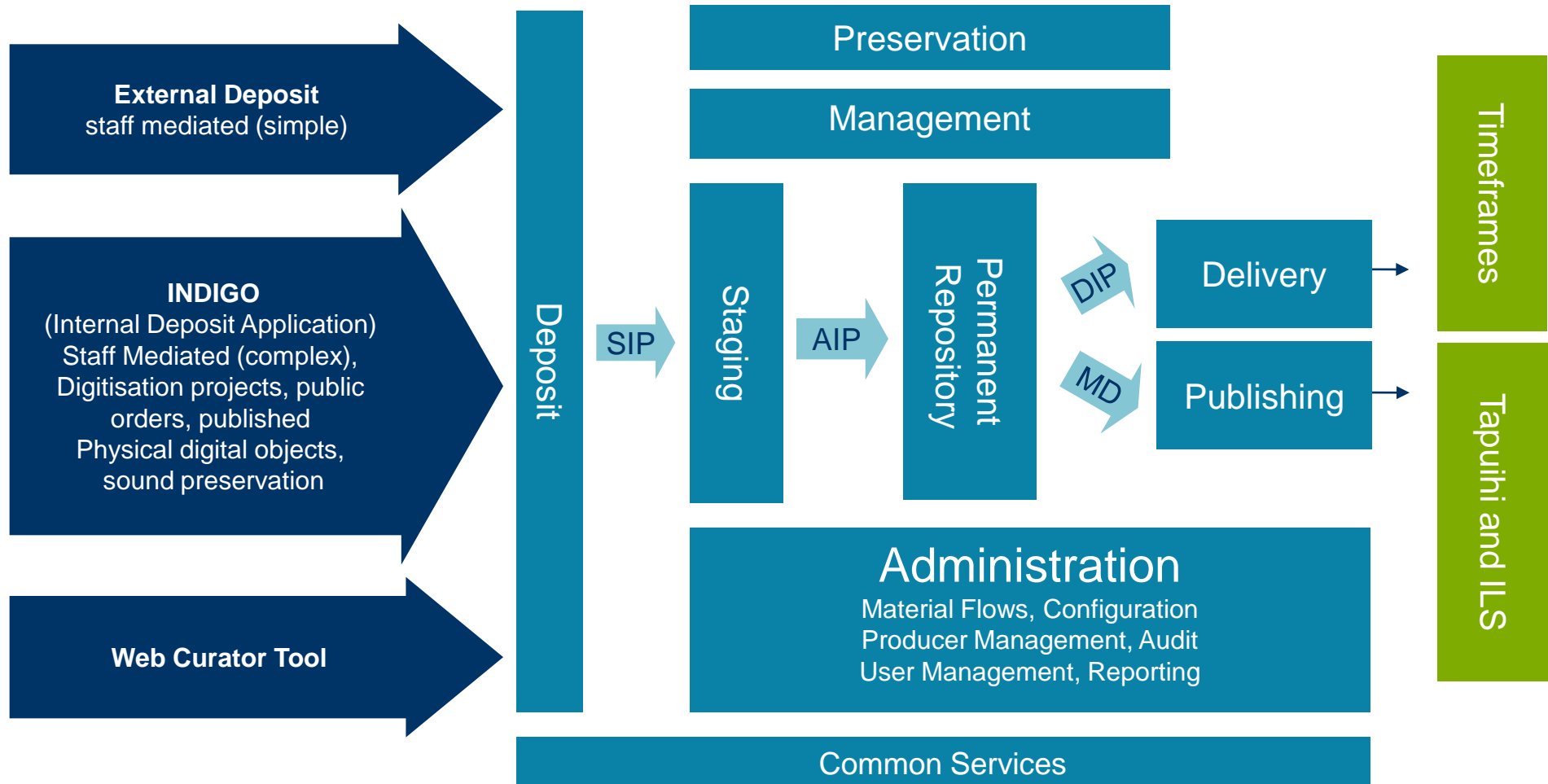


# Digital Preservation System





# The NDHA





# Some comments on obsolescence and risk management



# Assumptions/Assertions

'Managing format is  
fundamentally  
important'

Steve Abrams (iPres  
2008)

- We do not deny technology  
(complexity is not necessarily bad)
- We accept all formats
- We do not presume that our risk is the same as  
others' (and vice versa)
- We are keeping everything selected for ...(ever?)



# Some rules of engagement

Some attributes of a  
preservation risk  
assessment process

- Risk assessment has to be:
  - Automated (to a degree)
  - Meaningful and obtainable
  - Granular
  - Cognizant of internal and external factors
  - Able to be acted on...(bytestream)



# Risk *of* what?

Does rethinking our model from 'risk to' to 'risk of' help us in better framing a preservation risk assessment process?

## **“Obsolescence/obsolete/obsoleteness”**

Q1: What does obsolescence/obsolete mean?

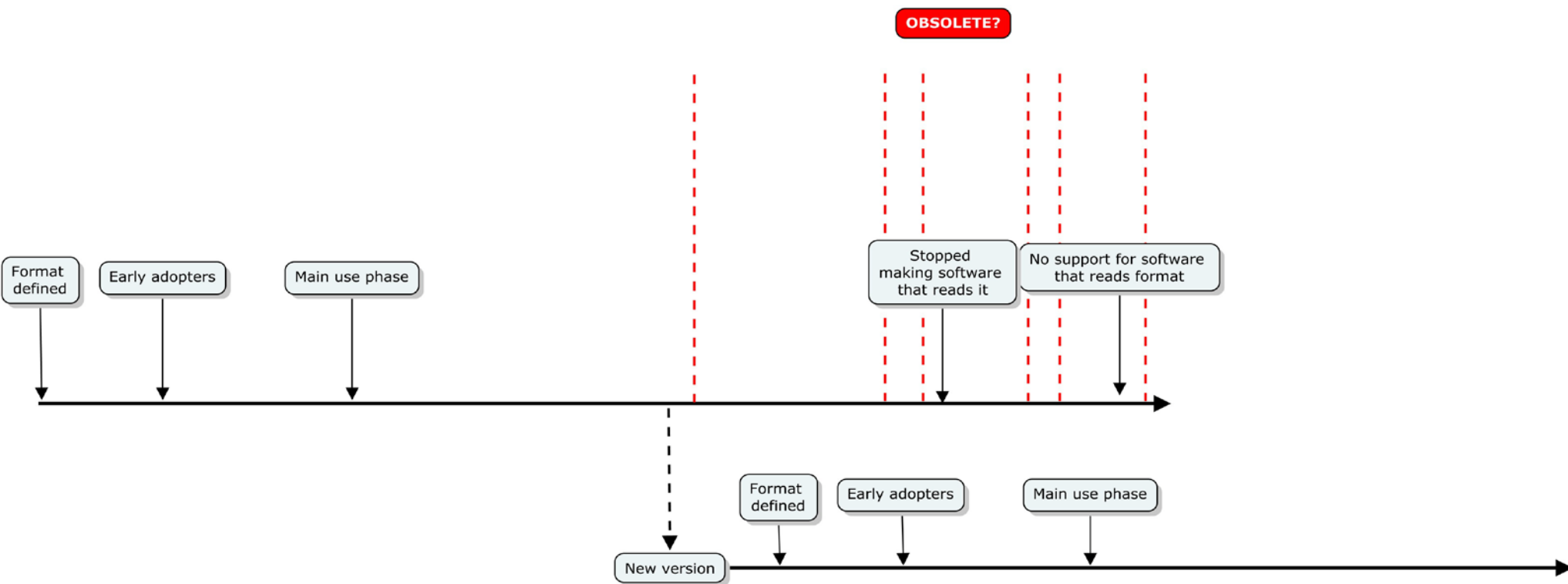
Q2: How do we recognise its approach?

Q3: How can we quantify this for analysis?



# Obsolescence

## Defining the Point of Obsolescence





# NLNZ definition of obsolescence

‘Risk is about the  
impending loss of the  
means of providing  
access’

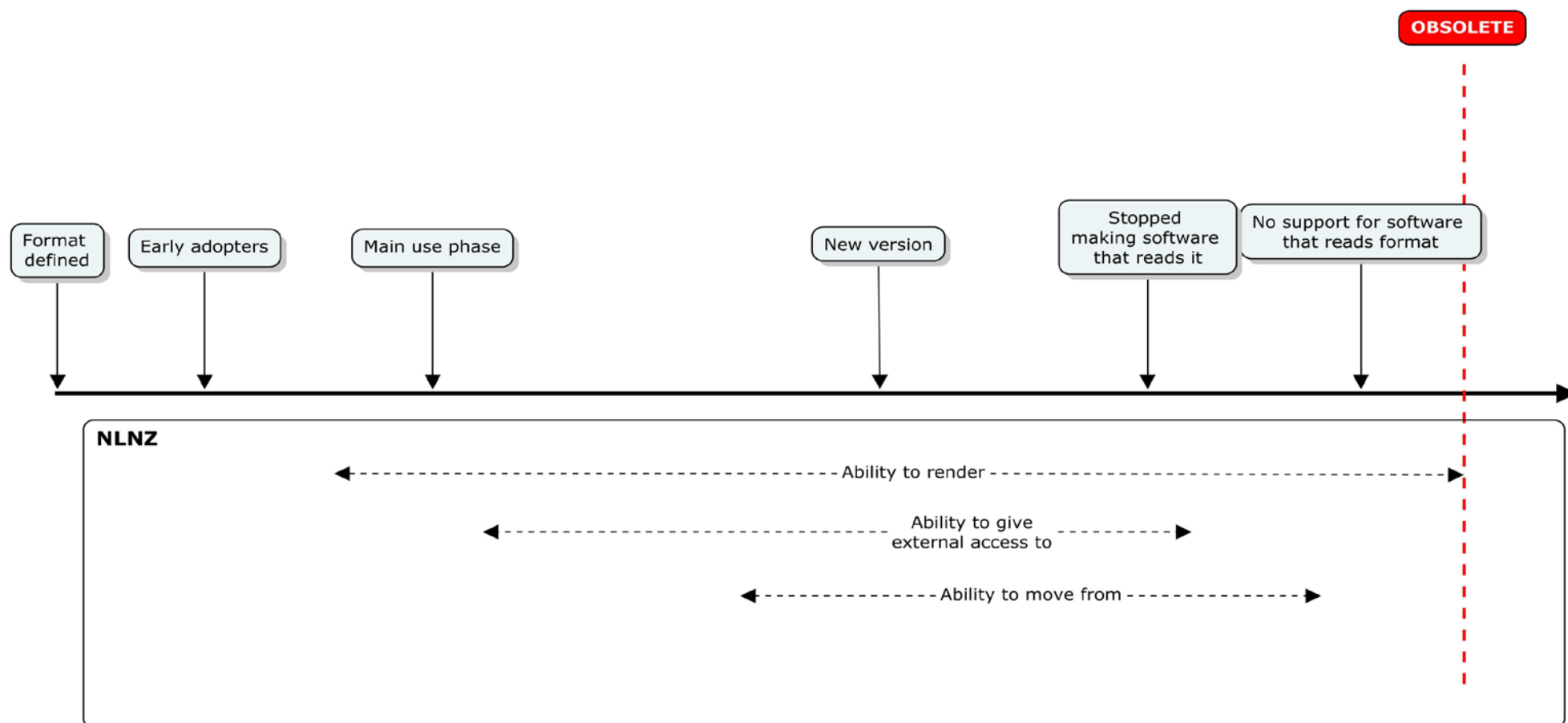
Pearson & Webb  
*IJDC 1:3, 2008*

- We define format obsolescence in relation to the Library’s ability to render files within the repository.
- If we cannot view, render, or migrate formats then they are “at-risk”.



# NLNZ point

## Defining the Point of Obsolescence





# Proposed Solution:

Here's the link with the idea of a Unified Digital Format Library (UDFR)

- Institutional Libraries that:
  1. Ensure the NDHA has a precise understanding of the contents of the permanent repository and **what degree of it can and cannot be rendered.**
- And then to:
  2. Have a warning system that gives “enough” time to take action to stop files becoming inaccessible.



# Library/Registry Components

A tripartite approach  
to risk assessment

- A Local Format Library
- An Application Library  
(that records the Library's available or tested tools)
- A Risk Library  
(that documents known problems that can affect our ability to render digital objects)



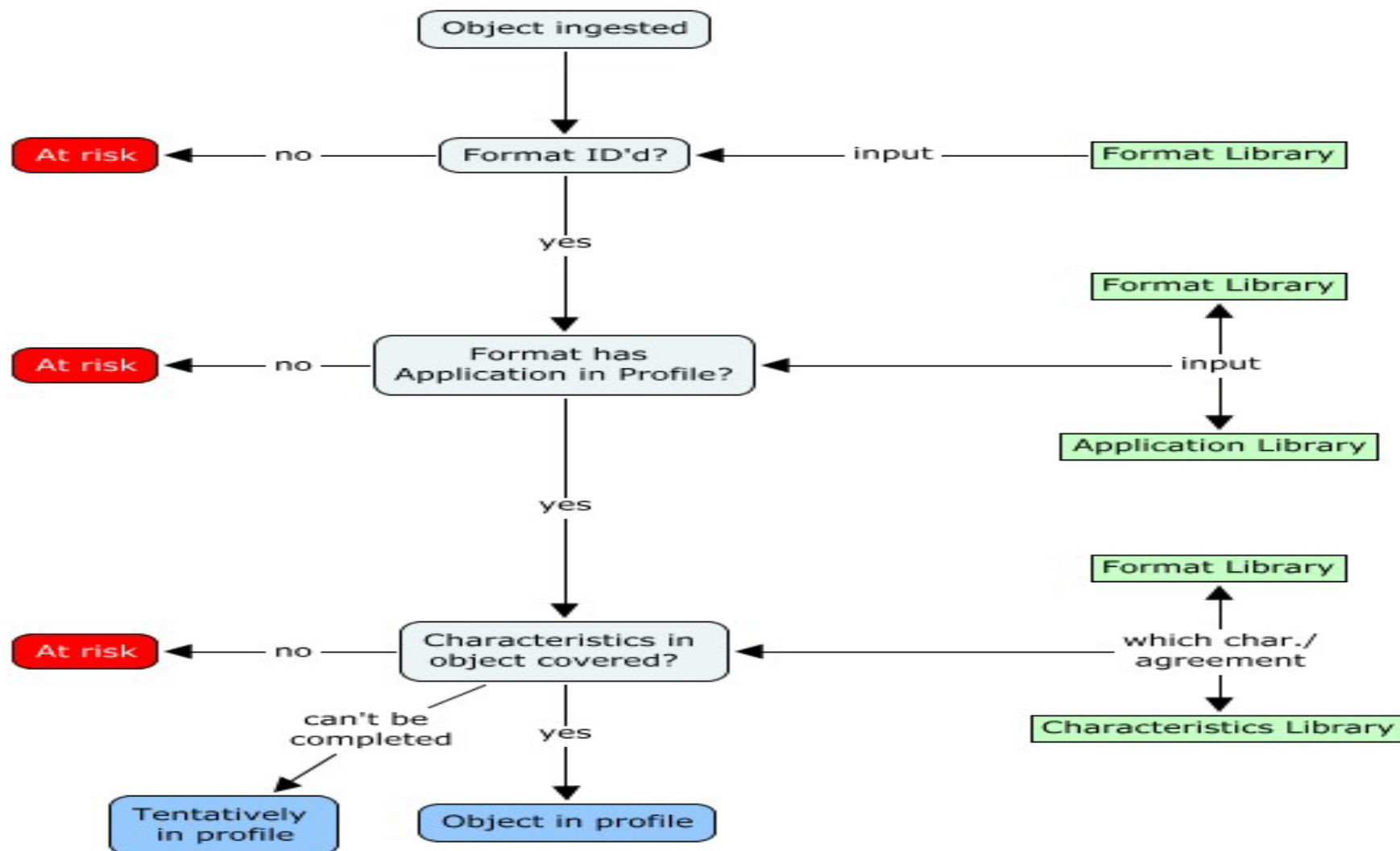
# The Libraries Will Document

Ok, so what sort of stuff is going to be in these three libraries?

- Formats that can be rendered;
- Specific versions of formats that can be rendered;
- The particular characteristics within these versions that are “problematic” (for example compression and colour encoding);
- Applications that can render variations of formats; version and characteristics;
- The sustainability of applications and formats.



# A Risk Management Decision Tree





# What else is in the Libraries?

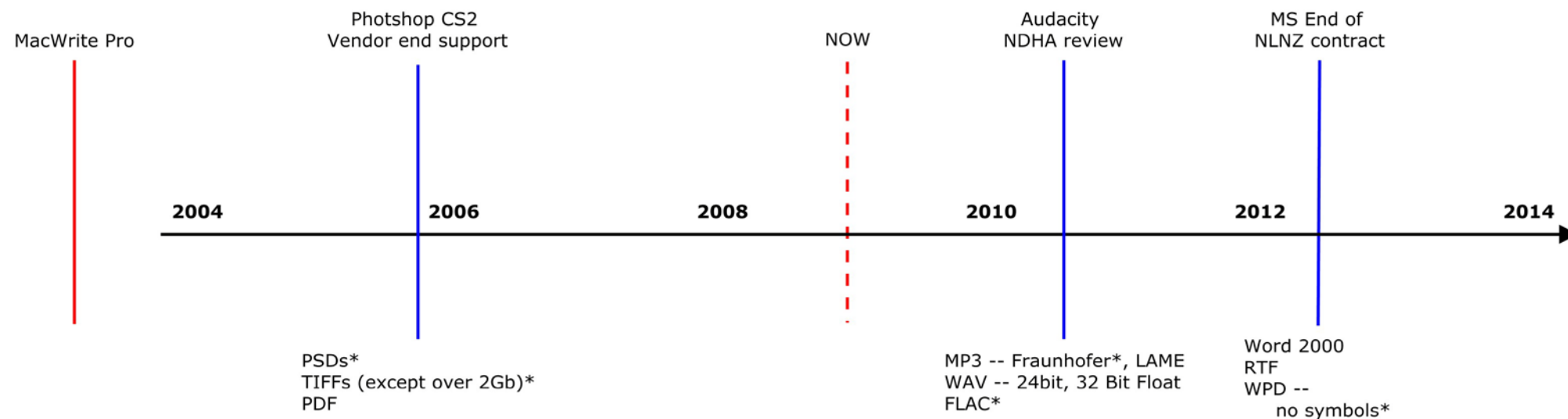
Defining the timelines  
related to particular  
format risk criteria

- Our Application library also tells us an amount of time we know/hope we will have the rendering application for, through:
  - a. Contract dates with vendor
  - b. Tech services schedules
  - c. Controlling the application in the system
  - d. Vendor support dates
  - e. Review date if no other date in place.



# Application timescales

## Application timescales





# Format library and risk grading

Some more on the link to UDFR and other format based tools and services such as PRONOM and DROID

- A local library of formats connected to the global registry.
- The Goal:
  - Extend the global with local information.
  - Extend the global with additional formats
- Works with PRONOM as a global registry.
- Connected to – application library, characteristics and risk.
- Each format can have one or more risks attached.
- A risk can refer to sub set of the format.
- Risks are updated by users and can be global or



# In Summary

- NLNZ risk management is based on capability: can we render it?
- This is a relationship between formats and applications
- Characteristics of formats create rendering issues
- We can control our own application destiny (sort of...)



# Where are we up to in digital preservation?



# Where are we up to in digital preservation?

What are some of the issues we face?

- Language
- The data deluge
- Products, tools and services
- Quality assurance and confidence
- Drivers towards standards/ best practice
- Economic sustainability
- Challenges ahead



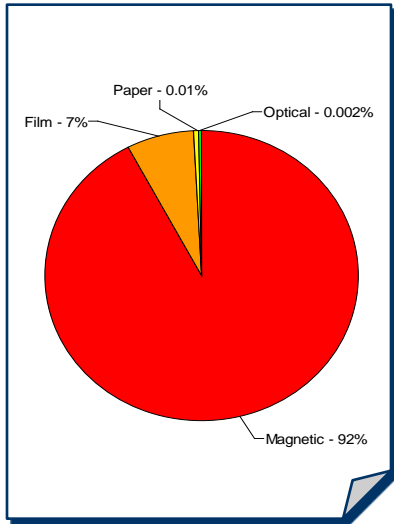
What do we mean  
when we talk about  
digital preservation?

- Repositories
- Data archiving
- Digital archiving
- Life cycle
- Digital curation
- Data curation
- Digital preservation

What does the language we use do to  
enlighten or confuse or obfuscate  
what we're trying to achieve in digital  
preservation?



# More about the data deluge



## BBC

Petabytes per week

## CERN LHC – black holes (mini or otherwise)

How much data?

## Content complexity

- Kam Woods – CDs
- Alex Ball – CAD (Engineering)
- Mark Guttenbrunner (gaming)
- Astronomy, oceanography
- Management of data sets

Work on digital preservation is really only just beginning.



# Digital preservation – risk management

‘Managing format is fundamentally important’

Steve Abrams (iPres 2008)

Concern about formats is concern about risk management

We need

- a comprehensive management approach
- a strategy that identifies the risk of format obsolescence
- a strategy that mitigates the risk of format obsolescence

This depends on

- our ability to identify the specific files that are most at-risk
- ready access to detailed, accurate information describing the file formats



# Products, tools and services

There would be significant benefit in standardising the tools we use for identification, validation and extraction

- Current tools performing format identification (and therefore assisting in risk identification) are JHOVE, DROID, and MET
- Limited formats, overlapping functionality
- No tool is currently capable of dealing with a range of formats in a satisfactory manner
- PRONOM, JHOVE and MET all deal with a very limited number of formats
- Problems regarding the accuracy of all are well documented
- If institutions don't define selection by format, development of new tools and modules for existing tools will continue to be needed



# Standards/best practice

Where's the agreement as to what comprises digital preservation?

- OAIS
  - PREMIS
  - NARA
  - PLANETS
  - NDIIPP
  - CASPAR
  - SHAMAN
  - DURASPACE
  - HathiTrust
- 
- Requirements
  - Certification
  - Audit



# Economic sustainability

Sustaining the digital  
investment ...

Blue Ribbon  
Task Force

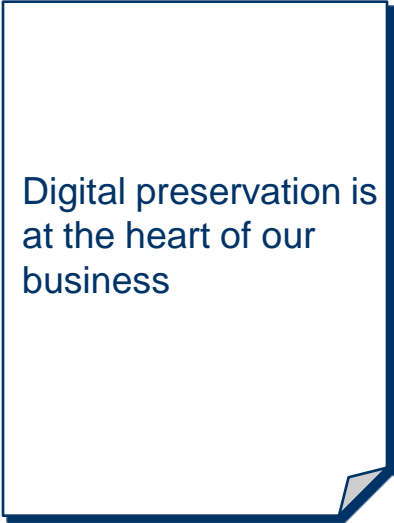
December 2008

‘In many institutions and enterprises systemic challenges create barriers for sustainable digital access and preservation’ including:

- Inadequacy of funding models to address long-term access and preservation needs
- Confusion and/or lack of alignment between stakeholders, roles, and responsibilities with respect to digital access and preservation
- Inadequate institutional, enterprise, and/or community incentives to support the collaboration needed to reinforce sustainable economic models
- Complacency that current practices are good enough
- Fear that digital access and preservation is too big to take on



# Challenges ahead



Digital preservation is  
at the heart of our  
business

- Agreed lexicon describing what we mean by digital preservation and what we want from digital preservation systems
- Capability/capacity to respond to technological change and innovation
- Citizen's created content impacting on our collection, description and preservation processes
- Content (ie digital preservation) systems are our core operational systems, not the catalogue
- Defining, resourcing and pursuing the research agenda (understanding the web, science data sets etc)
- Quality assurance of products and tools
- Professional services market (commercial or otherwise)
- Digital preservation as a component of a national knowledge infrastructure
- A coordinated national/international approach to supporting digital preservation research, products and services



Thank you

[Steve.knight@natlib.govt.nz](mailto:Steve.knight@natlib.govt.nz)