e●permanence

The National Archives of Australia: Approach to digital preservation

Andrew Wilson & Cornel Platzer
National Archives of Australia:
Digital Preservation Project
15 September 2004

Topics

- concepts (Andrew)
- implementation (Cornel)
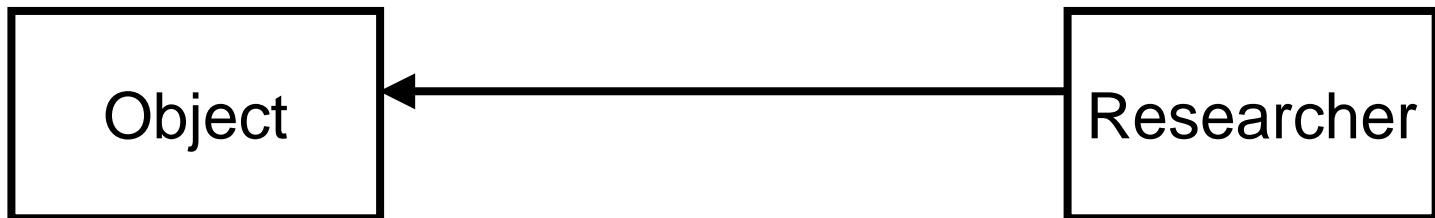- XENA demo (Andrew)

Concepts

Essential performance model

- idea of 'essential performance' is central to the NAA approach

- two significant components to this:

  - **performance**

  - **essence**

Traditionally

- records are experienced directly as objects

```
┌─────────────┐                      ┌──────────────┐
│             │ ◄─────────────────── │              │
│  Object     │                      │  Researcher  │
│             │                      │              │
└─────────────┘                      └──────────────┘
```

- preserving the object is how we preserve the 'record'

Electronic records are different

- 'record' is experienced as a performance

```
┌──────────┐     ┌──────────┐     ┌──────────────┐     ┌──────────────┐
│  Object  │ ──► │ Process  │ ──► │ Performance  │ ◄── │  Researcher  │
└──────────┘     └──────────┘     └──────────────┘     └──────────────┘
```
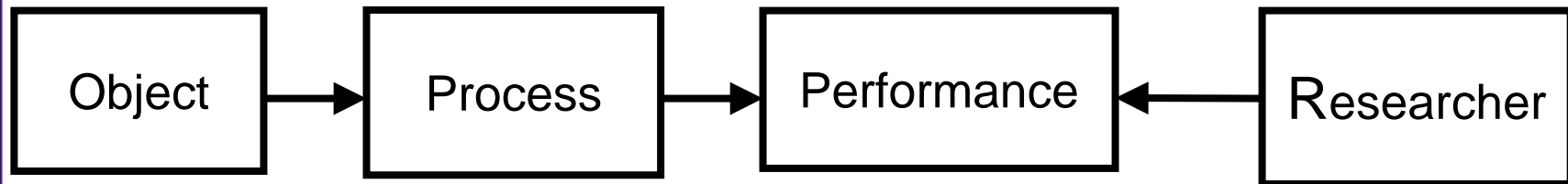
- preserve the performance and we preserve the 'record'

Example: email message

Source	Process	Performance
Data File (sequence of 0s and 1s)	Outlook Email Client Windows 2000 Intel Computer	Bitmaps on Monitor Paper Print-Out

Researcher
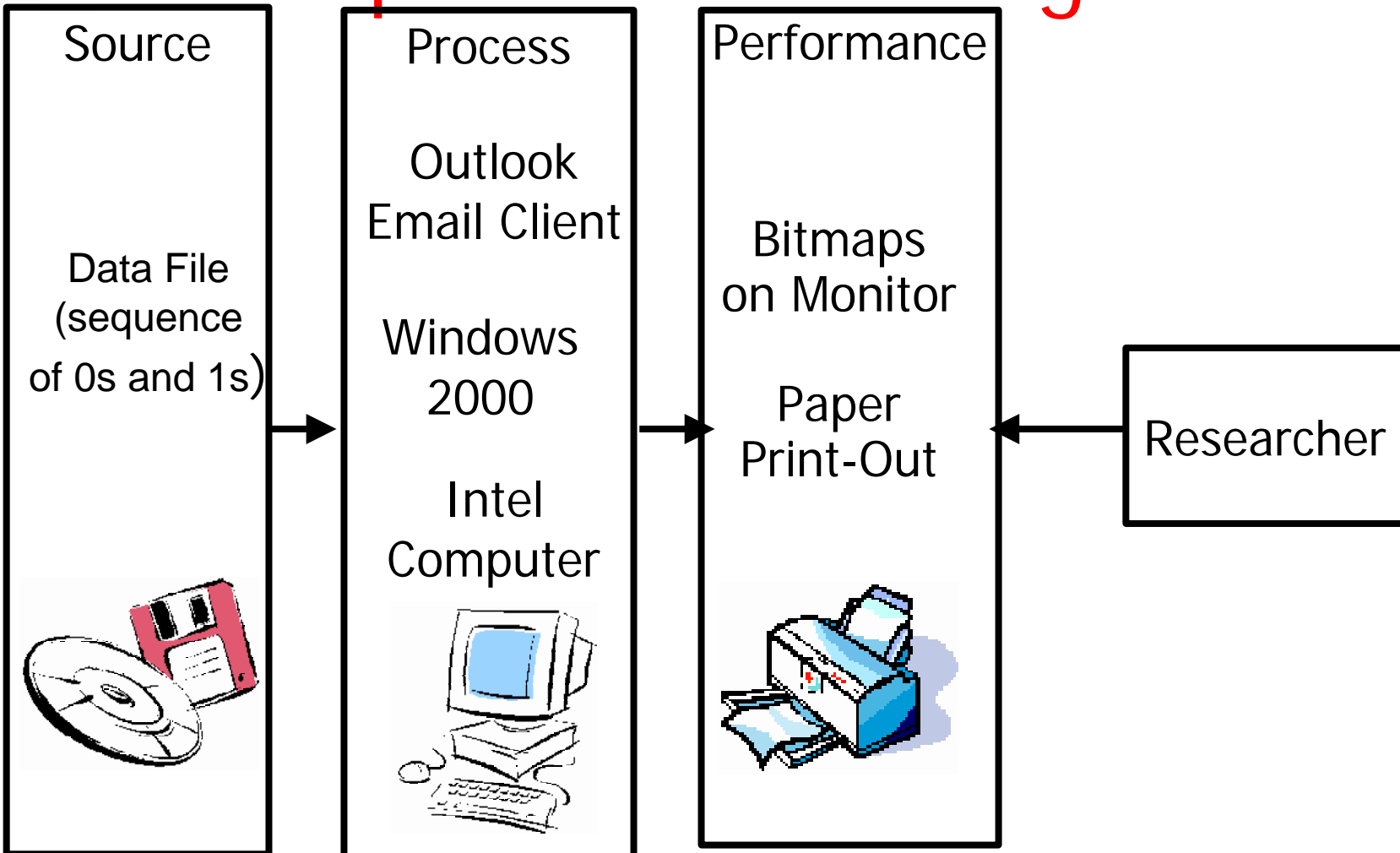
Keeping source + process doesn't work

- media is fragile and fails
  - need to refresh the source

- components become obsolete
  - technological and marketplace obsolescence

- components are proprietary and licensed (not purchased)
  - software licenses don't last forever!

so…we need a different approach

but what?

Components of the approach

1. what is significant about the performance – what is the *essence*?

2. what format best expresses the *essence*?

3. what do we use to manage the new source (*essence*) in a way that allows us to reproduce the performance at will?

Preserving performances

```
  ( Acquire        ( Replace )      ( What is
    new                             significant? )
    format )
      |                 |                |
      |                 |                |
      v                 v                v
 ┌─────────┐     ┌─────────┐     ┌─────────────┐     ┌────────────┐
 │ Object  │ ──▶ │ Process │ ──▶ │ Performance │ ◀── │ Researcher │
 └─────────┘     └─────────┘     └─────────────┘     └────────────┘
```

1. What is significant?

Object → Process → **Performance** ← Researcher
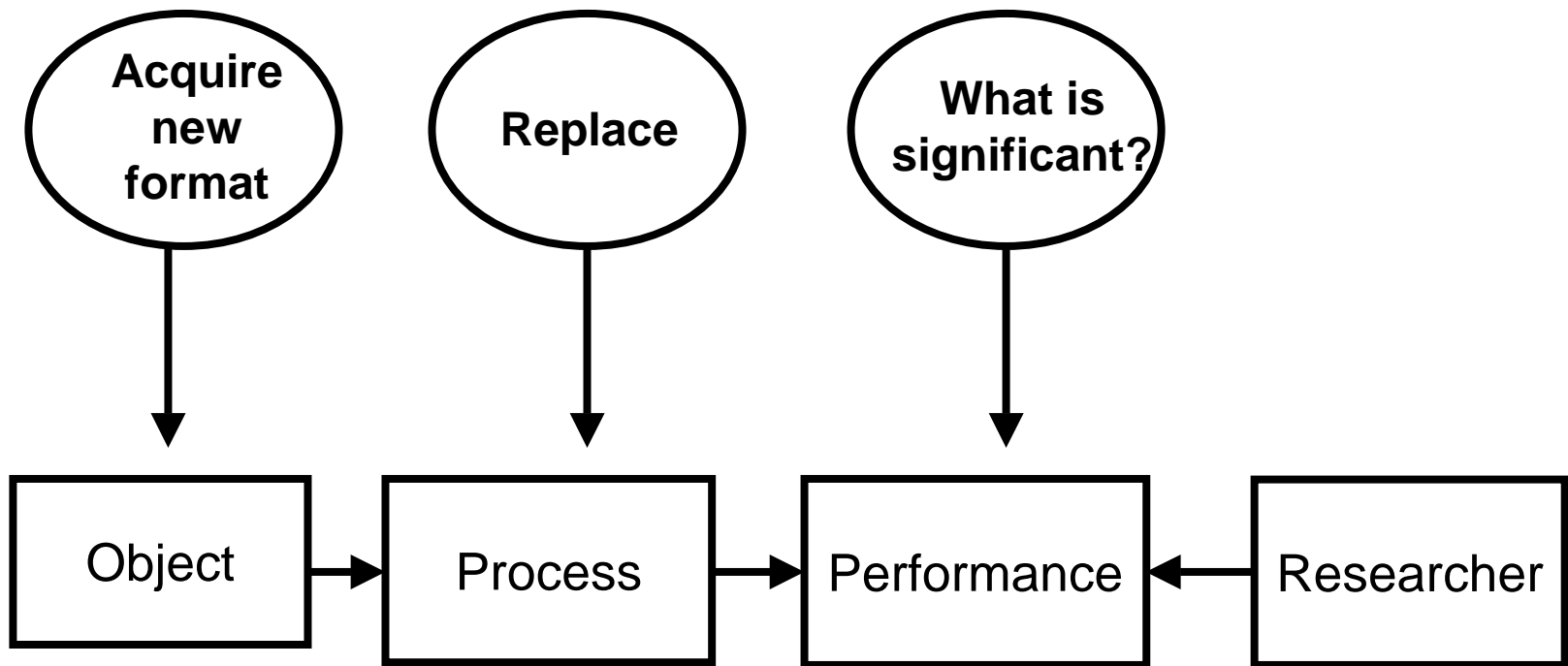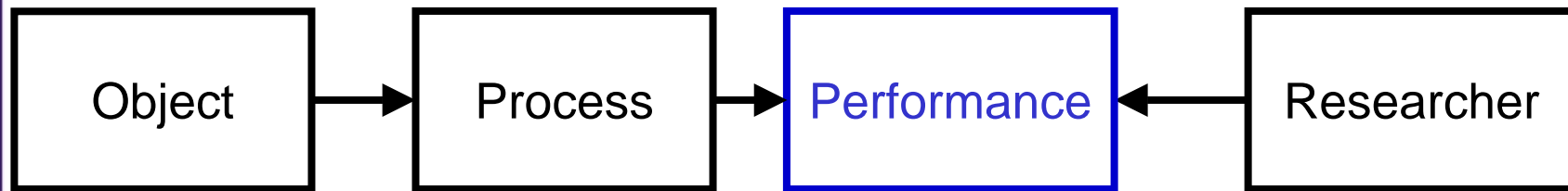
Determine the essential aspects of the performance

The essence of the record

The '*essence*' of the record

- what we want to preserve out of the performance
  - what aspects are essential to the record's value?
  - what aspects are incidental to the record's value?

- consider census records
  - microfilm copies capture the essence of the census return so no need to keep originals

APPENDIX

I. Digital Preservation Principles

1. *The digital preservation program must be able to preserve any digital record that is brought into National Archives' custody regardless of the application or system it is from or data format it is stored in.*

2. *The digital preservation program must determine and preserve the essence of the digital records in the National Archives' custody and recreate their essential performance over time.*

3. *The digital preservation program will be based on non-proprietary technologies.*

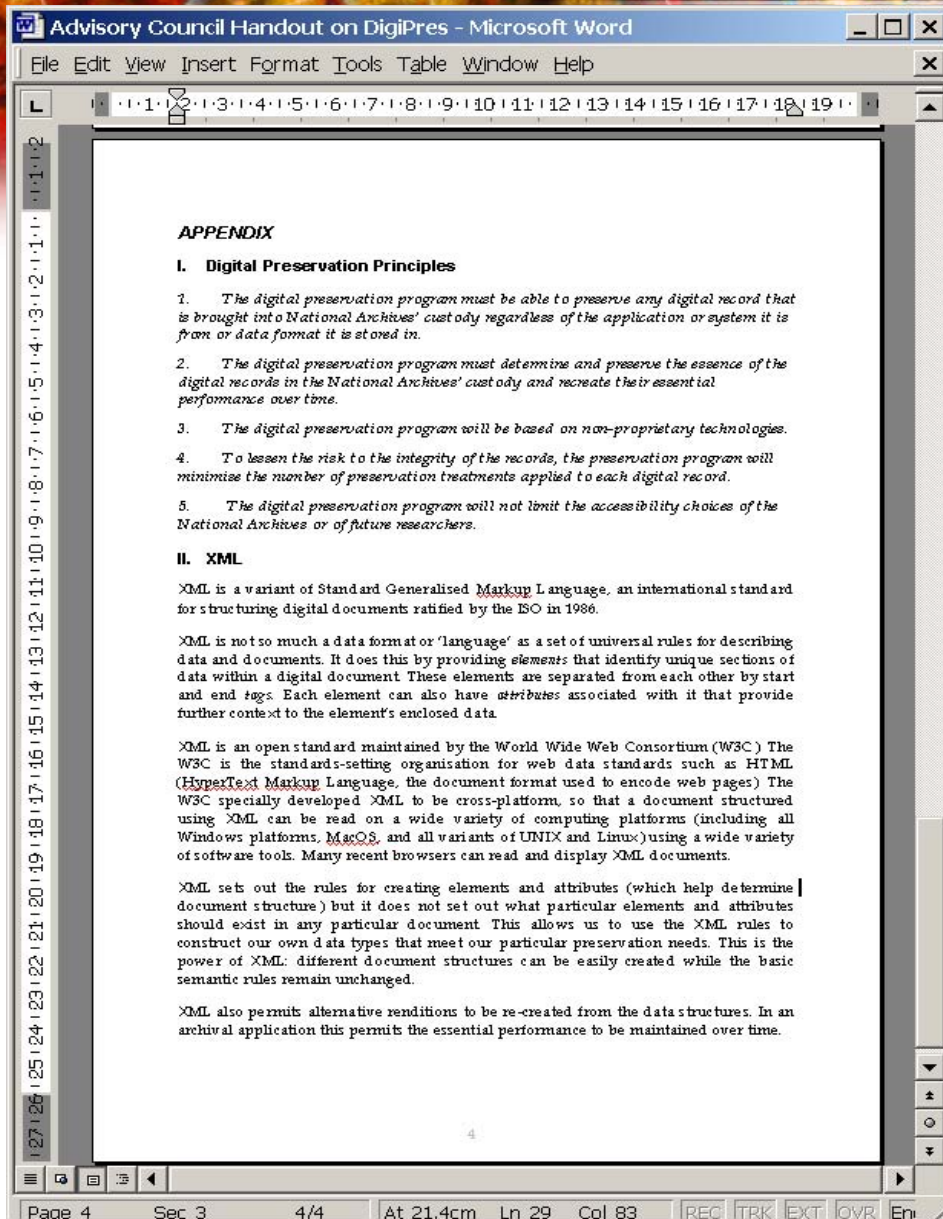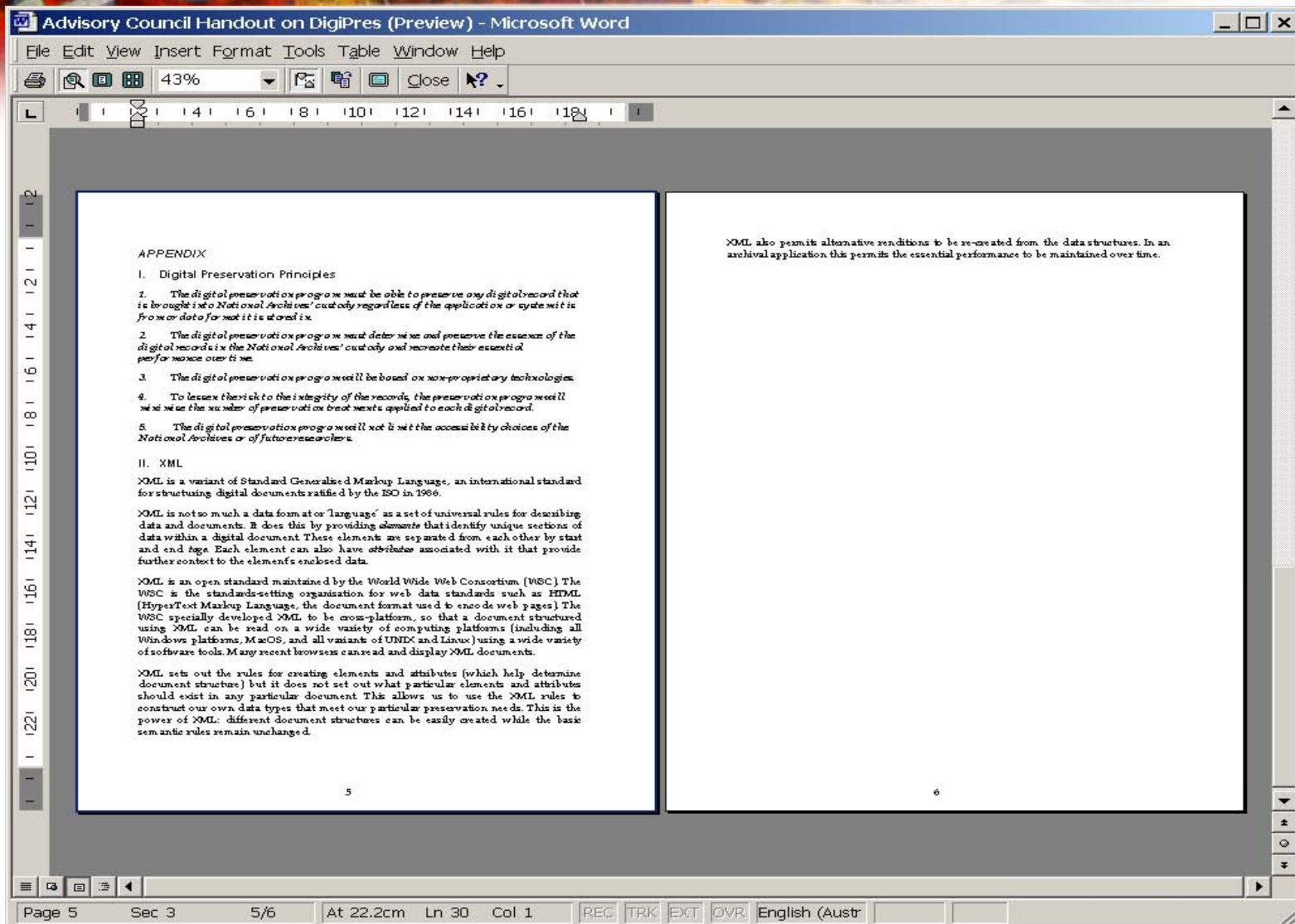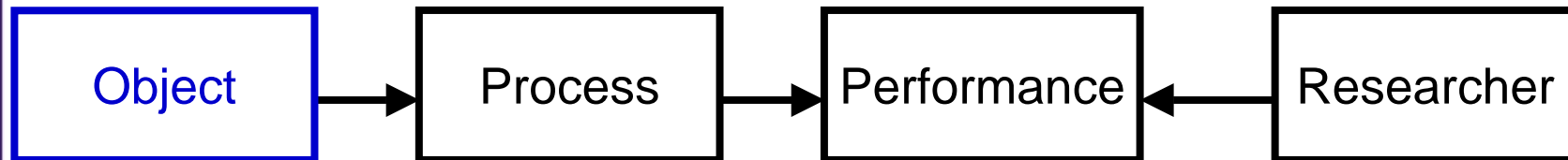4. *To lessen the risk to the integrity of the records, the preservation program will minimise the number of preservation treatments applied to each digital record.*

5. *The digital preservation program will not limit the accessibility choices of the National Archives or of future researchers.*

II. XML

XML is a variant of Standard Generalised Markup Language, an international standard for structuring digital documents ratified by the ISO in 1986.

XML is not so much a data format or 'language' as a set of universal rules for describing data and documents. It does this by providing *elements* that identify unique sections of data within a digital document. These elements are separated from each other by start and end *tags*. Each element can also have *attributes* associated with it that provide further context to the element's enclosed data.

XML is an open standard maintained by the World Wide Web Consortium (W3C). The W3C is the standards-setting organisation for web data standards such as HTML (HyperText Markup Language, the document format used to encode web pages). The W3C specially developed XML to be cross-platform, so that a document structured using XML can be read on a wide variety of computing platforms (including all Windows platforms, MacOS, and all variants of UNIX and Linux) using a wide variety of software tools. Many recent browsers can read and display XML documents.

XML sets out the rules for creating elements and attributes (which help determine document structure) but it does not set out what particular elements and attributes should exist in any particular document. This allows us to use the XML rules to construct our own data types that meet our particular preservation needs. This is the power of XML: different document structures can be easily created while the basic semantic rules remain unchanged.

XML also permits alternative renditions to be re-created from the data structures. In an archival application this permits the essential performance to be maintained over time.

4

NATIONAL ARCHIVES OF AUSTRALIA

APPENDIX

I. Digital Preservation Principles

1. The digital preservation program must be able to preserve any digital record that is brought into National Archives' custody regardless of the application or system it is from or data format it is stored in.

2. The digital preservation program must determine and preserve the essence of the digital records in the National Archives' custody and recreate their essential performance over time.

3. The digital preservation program will be based on non-proprietary technologies.

4. To lessen the risk to the integrity of the records, the preservation program will minimise the number of preservation treatments applied to each digital record.

5. The digital preservation program will not limit the accessibility choices of the National Archives or of future researchers.

II. XML

XML is a variant of Standard Generalised Markup Language, an international standard for structuring digital documents ratified by the ISO in 1986.

XML is not so much a data format or 'language' as a set of universal rules for describing data and documents. It does this by providing *elements* that identify unique sections of data within a digital document. These elements are separated from each other by start and end *tags*. Each element can also have *attributes* associated with it that provide further context to the element's enclosed data.

XML is an open standard maintained by the World Wide Web Consortium (W3C). The W3C is the standards-setting organisation for web data standards such as HTML (HyperText Markup Language, the document format used to encode web pages). The W3C specially developed XML to be cross-platform, so that a document structured using XML can be read on a wide variety of computing platforms (including all Windows platforms, MacOS, and all variants of UNIX and Linux) using a wide variety of software tools. Many recent browsers can read and display XML documents.

XML sets out the rules for creating elements and attributes (which help determine document structure) but it does not set out what particular elements and attributes should exist in any particular document. This allows us to use the XML rules to construct our own data types that meet our particular preservation needs. This is the power of XML: different document structures can be easily created while the basic semantic rules remain unchanged.

5

XML also permits alternative renditions to be re-created from the data structures. In an archival application this permits the essential performance to be maintained over time.

6

NATIONAL ARCHIVES OF AUSTRALIA

2.   What format?

```
┌──────────┐     ┌──────────┐     ┌──────────────┐     ┌──────────────┐
│  Object  │ ──▶ │ Process  │ ──▶ │ Performance  │ ◀── │  Researcher  │
└──────────┘     └──────────┘     └──────────────┘     └──────────────┘
```

Normalise into open, fully documented
data formats that can represent that essence
Formats based on XML

Normalisation

- don't try to preserve the original process at all
- develop our own processes using open public-domain standards
- commit to maintain those processes over time
- one-off conversion of the source to suit *our* processes
- we set our own preservation platform
  - Not reacting to others
- aids authenticity by not changing the master regularly

3.   What tool?

Object → Process → Performance ← Researcher
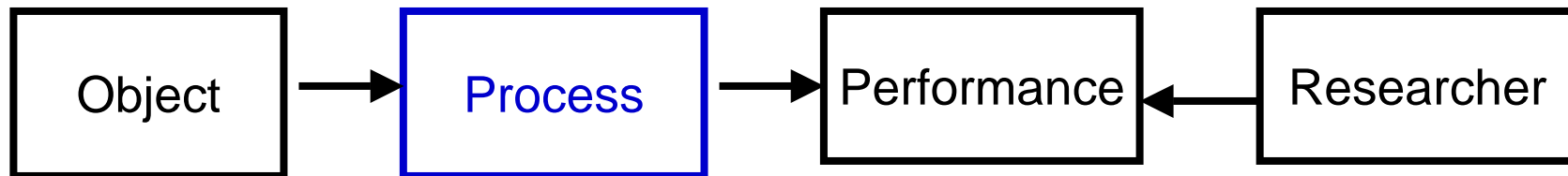
Acquire technologies to manage those formats

Our tool: Xena

Xena

- our tool for preserving the 'essential performances'
- it normalises data objects and…
- …it re-presents performances
- written in Java
  - Windows, Linux, Mac OS X
- open under the GPL
- latest at: sourceforge.net/projects/xena/

III   XENA

(**x**ml **e**lectronic **n**ormalising of **a**rchives)

XENA core

The core consists of:

    I. graphical user interface components

    II. plug-in management components

    III. generic validation components

XENA plug-ins

Plug-ins are created for identified data types that are to be processed. Each plug-in includes:

- guesser component
- one or more input format type components
- normalised format type (xml schema)
- one or more normalisation modules
- one or more view components
- validation functionality

A 'xena'd' object

Metadata

Content

Checksum
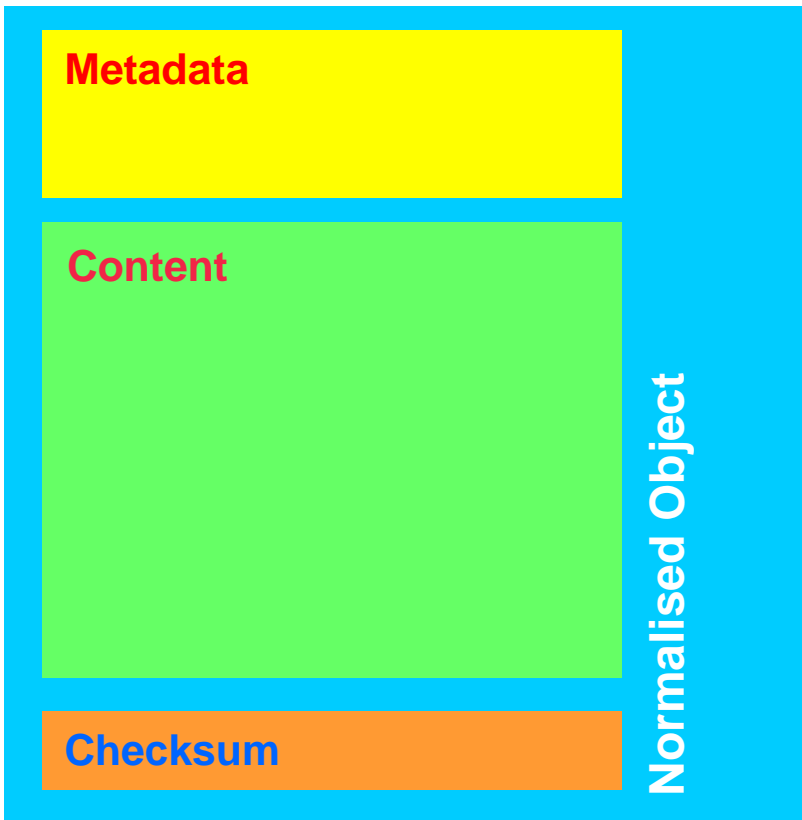
Normalised Object

What the Xena
normalisation
process produces: a
wrapped object

And now the demonstration …