

Digital Preservation: Preserving the Past

Peter Lloyd

Tessella Support Services

26 October 2006



Contents

- What's the problem?
- Issues.
- What is a digital record?
- Approaches to preservation.
- Solution frameworks.
- Real projects.

Digital Preservation Mission

Taken from the United States NARA ERA project:
To preserve any type of electronic record,
created using any type of application,
on any computing platform,
delivered on any digital media,
from any entity in the Federal Government and
any donor.



What ' s the Problem?

- 1986 Domesday Project
 - Original Domesday book, produced in 1086, still readable today.
 - BBC 900th anniversary project.
 - Used state-of-the-art technology.
 - 30 cm laser discs
 - BBC Microcomputers



What ' s the Problem?

- Morgan Stanley

- Court requested emails dating back to 1998.
- E-mails held on back-up tapes.
- June 2004 Morgan Stanley certified that all documents had been handed over.
- "The storage folks found an additional 1,600 backup tapes in a closet,"



- \$1 Billion damages awarded against them



What's the Problem?

- Viking Lander data
 - Mission data stored since 1975 on magnetic tape.
 - Stored in climate-controlled environment.
 - Tapes and data formats unreadable 20 years later.
 - Retype everything!

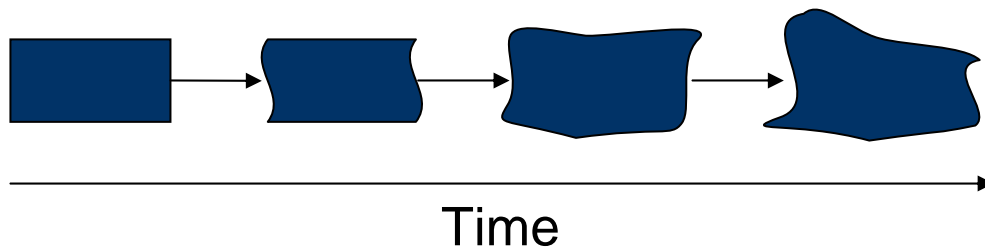


What's the Problem?

- Need to keep data for:
 - Legal requirements
 - Accountability
 - Shaping the long-term view
 - Protecting data assets
 - Historical significance
 - Data mining
 - Business efficiency

Digital Preservation Issues

- Cannot be read directly:
 - Rely on file format
 - Rely on application software
 - Rely on O/S
 - Rely on storage media
 - Rely on hardware
- Must ensure *loss/less* transformation



Digital Preservation Issues

- Volume of records
 - 5 exabytes of new information in 2002.
 - 92% stored on magnetic media.





(Source: <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003>)



What is a Digital Record?

- Electronic records characteristics:
 - Context Metadata
 - Content Format
 - Structure:Physical File hierarchy
 - Structure:Conceptual Format
 - Appearance Format/software
 - Behaviour Software

What is a Digital Record?

- Records of the Committee
 - Sub-committee 1
 - Agendas
 - Agenda 1 
 - Agenda 2 
 - ...
 - Minutes
 - Minutes 1 
 - Minutes 2 
 - ...
 - Sub-committee 2
 - ...

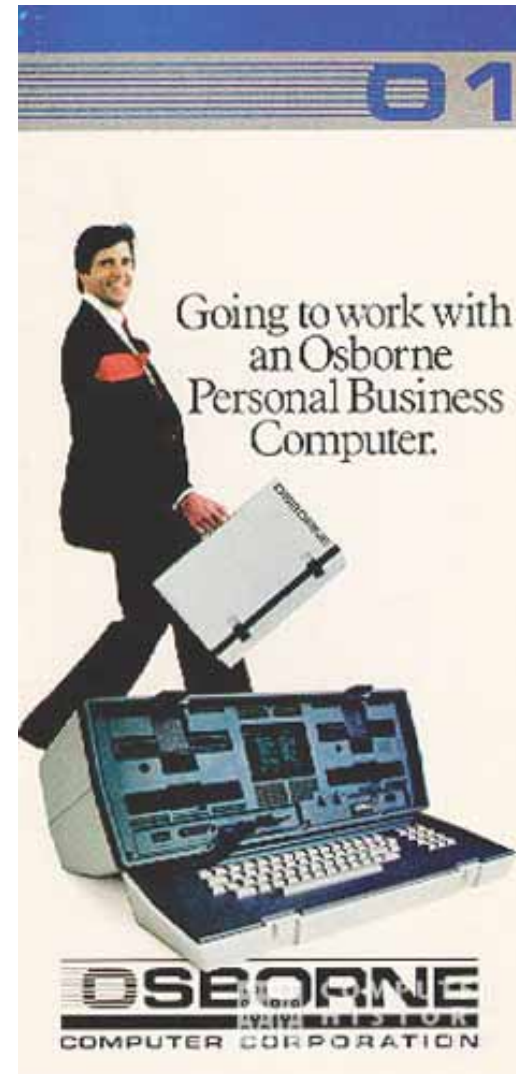
Preservation Approaches

- Hard Copy
 - Technically simple.
 - Lose dynamic nature of digital records.
 - Impractical.



Preservation Approaches

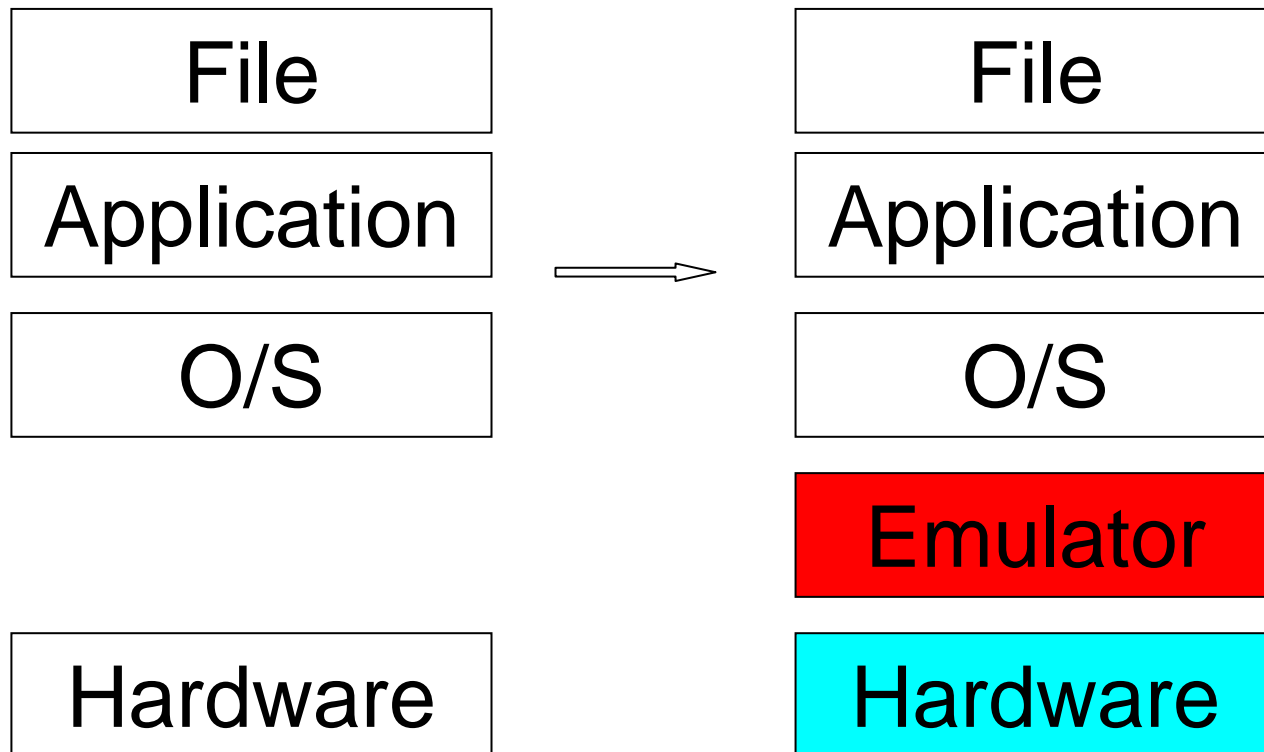
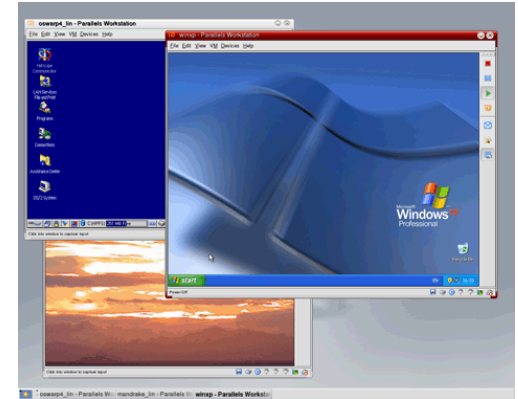
- Museum
 - Maintain old software/hardware in working order.
 - Need to consider all combinations.
 - Increasingly expensive.
 - May be used as interim measure.



Preservation Approaches

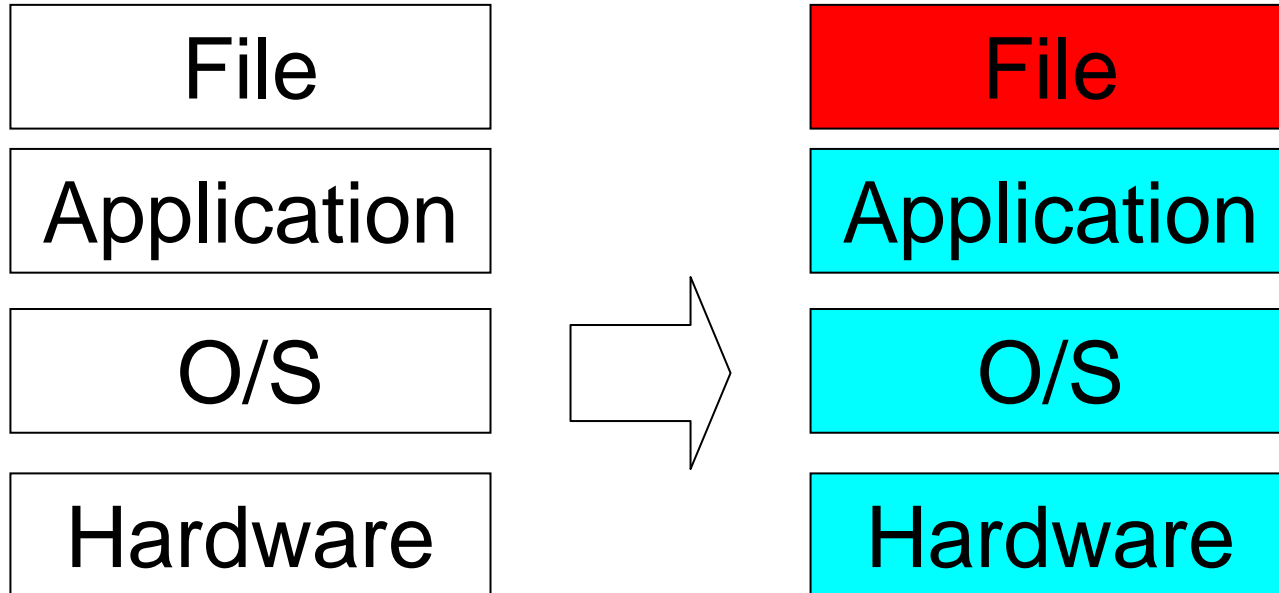
- Emulation

- Good for preserving behaviour
- Unproven as yet in practice
- Large software maintenance overhead



Preservation Approaches

- Migration



Preservation Approaches

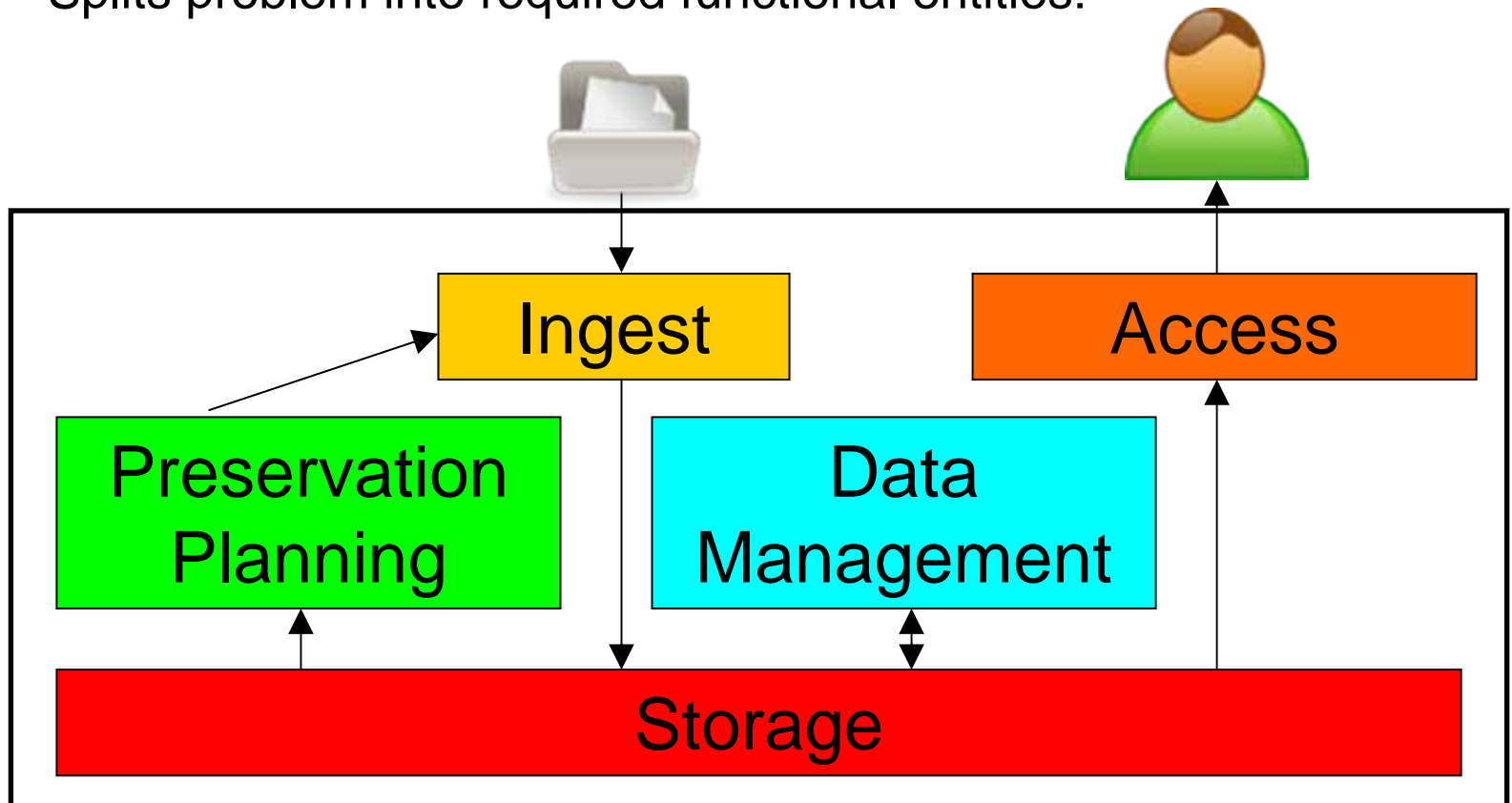
- Migration

- Metadata or content can be transformed to XML

```
<?xml version="1.0" encoding="UTF-8"?>
<record>
  <created>16-Oct-1980</created>
  <creator>Bob Smith</creator>
  <data>
    <value>0.76</value>
    <value>0.42</value>
  </data>
</record>
```

Solution Framework (OAIS)

- Framework come out of NASA.
- Not an application and doesn't tell you how to do it!
- Splits problem into required functional entities:



OAIS Segments

- **Ingest**
 - Record selection
 - Record structure
 - Metadata capture
- **Data management**
 - Change control and audits
- **Storage**
 - Metadata and Content
- **Access**
 - Finding and Disseminating records
- **Preservation planning**
 - Migration, Emulation, etc.
- **Administration**
 - Future proofing

Software Issues

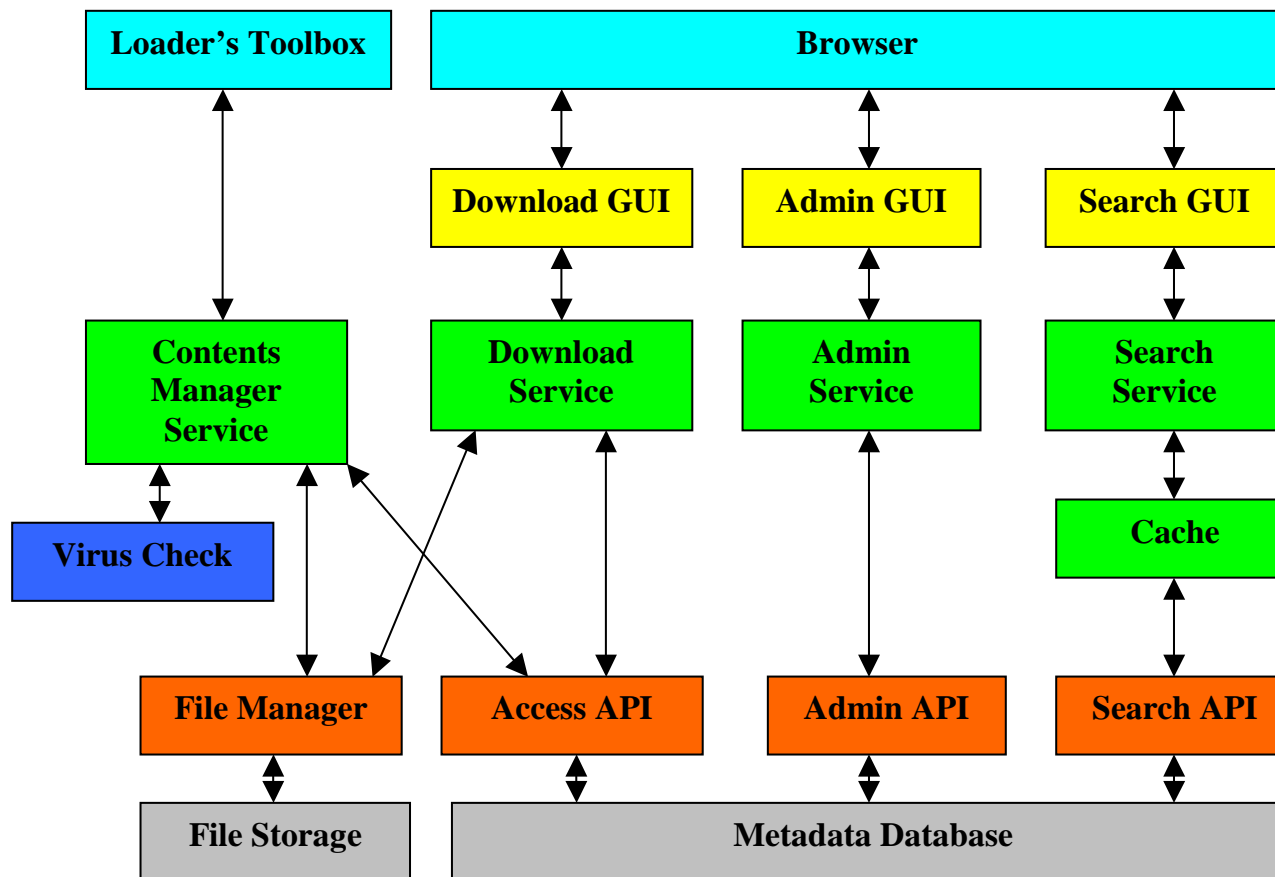
- Development technology
 - J2EE
 - .Net Framework
- Modular design
 - Clearly defined component interfaces
- Third-party components
 - 'Sunset' policy
- Operating systems
 - Avoid o/s dependant software if possible
- Hardware
 - Plan migrations well in advance

UK National Archives: Digital Archive



- Solution based on OAIS
- Open standards:
 - J2EE
 - Oracle 9i application server
 - XML/SOAP/JDBC
- e-GIF compliant.
- 100 TB of data in five years.
- Stores
 - Emails
 - Web-site snapshots
 - Images
 - Video clips
 - Sound files

UK National Archives: Digital Archive



UK National Archives: PRONOM

PRONOM | Search by format - Microsoft Internet Explorer

File Edit View Favorites Tools Help

> [Contact us](#) > [Help](#) > [A to Z index](#) > [Site search](#)

Tuesday 24 October

[Home](#) | [About us](#) | [Visit us](#) | [Research, education & online exhibitions](#) | [Search the archives](#) | [Services for professionals](#) | [News](#) | [Shop online](#)

You are here: [Home](#) > [Services for professionals](#) > [Preservation](#) > [PRONOM](#) > Search: Search by format

The technical registry
PRONOM

Welcome : About Add an entry
Search ? Help Information resources

Search : By format [? Help : search by format](#)

File format PRONOM Unique Identifier Software Vendor Lifecycles

1. File formats

Enter a file extension and click 'search' to find all file formats with that extension. Leave the file extension blank to find all file formats in the database.

* [Search >](#)

To search for a particular file format, enter the name of the file format and then click 'search'. Leave the file format name blank to find all file formats in the database.

[Search >](#)

2. Compatible software

Enter a file extension and click 'search' to find all software which can process in any way files with that extension.

* [Search >](#)

Enter a file format name and click 'search' to find all software which can process in any way files of that format.

[Search >](#)

[Terms of Use](#) | [Copyright](#) | [Privacy](#) | [Top of page](#)

The National Archives, Kew, Richmond, Surrey, TW9 4DU | email: enquiry@nationalarchives.gov.uk | tel: +44 (0) 20 8876 3444

PRONOM - File and Product Report Generator

start Peter Lloyd Mail - Mic... Tessella Forums - Mic... PRONOM | Search by... My Documents BCS Digital Preservati... Microsoft PowerP... 18:36

UK National Archives: PRONOM

The screenshot shows the PRONOM website interface. At the top, there is a navigation bar with links for Home, About, Visit us, Research, education & online exhibitions, Search the archives, Services for professionals, News, and Shop online. The date Tuesday 24 October is displayed on the right. Below the navigation bar, the breadcrumb trail reads: Home > Services for professionals > Preservation > PRONOM > Search by format > Results. The main heading is "The technical registry PRONOM". A search bar contains the text "Search Results" and a link to "Help: report on file format". Below the search bar, there are tabs for "File format", "PRONOM Unique Identifier", "Software", "Vendor", and "Lifecycles". The search results are displayed in a table with the following columns: PRONOM Unique ID, Format Name, Format Version, and Extension. The results show eight entries for the Portable Document Format (PDF) with versions 1.0 through 1.6, and one entry for the Portable Document Format - Archival (version 1). The page number "Page: 1 of 1" is shown at the bottom left of the results area. The browser's taskbar at the bottom shows several open applications, including Microsoft Internet Explorer, Peter Lloyd Mail, Tesella Forums, and Microsoft PowerPoint.

http://www.nationalarchives.gov.uk - PRONOM | Search by format - Microsoft Internet Explorer

File Edit View Favorites Tools Help

A the national archives

Contact us Help A to Z index Site search

Tuesday 24 October

Home About Visit us Research, education & online exhibitions Search the archives Services for professionals News Shop online

You are here: Home > Services for professionals > Preservation > PRONOM > Search by format > Results

The technical registry
PRONOM

Welcome : About Add an entry
Search ? Help Information resources

? Help : report on file format

Search Results

File format PRONOM Unique Identifier Software Vendor Lifecycles

You searched for: "pdf"

Save as... XML | CSV Print page 1

PRONOM Unique ID	Format Name	Format Version	Extension
fnt/14	Portable Document Format	1.0	pdf
fnt/15	Portable Document Format	1.1	pdf
fnt/16	Portable Document Format	1.2	pdf
fnt/17	Portable Document Format	1.3	pdf
fnt/18	Portable Document Format	1.4	pdf
fnt/19	Portable Document Format	1.5	pdf
fnt/20	Portable Document Format	1.6	pdf
fnt/95	Portable Document Format - Archival	1	pdf

Page: 1 of 1 page 1

PRONOM - File and Product Report Generator

start Peter Lloyd Mail - Mic... Tesella Forums - Mic... http://www.nationals... My Documents BCS Digital Preservati... Microsoft PowerP... 18:42

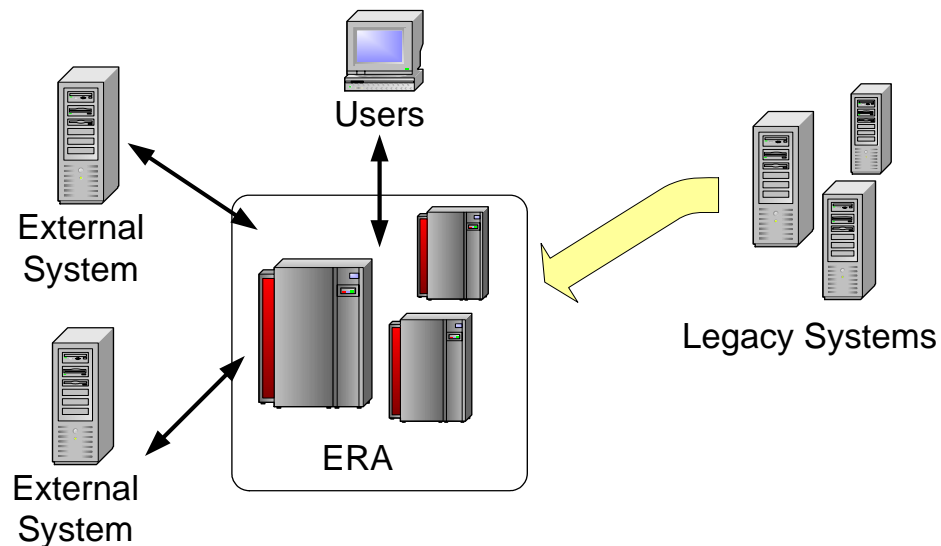
Dutch Government: Digital Preservation Testbed

nationaal archief

- Research project to evaluate digital preservation approaches.
- Structured as a series of experiments.
- Preservation approaches:
 - Migration
 - Emulation
 - Conversion to standard formats (XML)
 - Combinations
- Record types:
 - Documents (e.g. MS Word)
 - E-mails
 - Databases

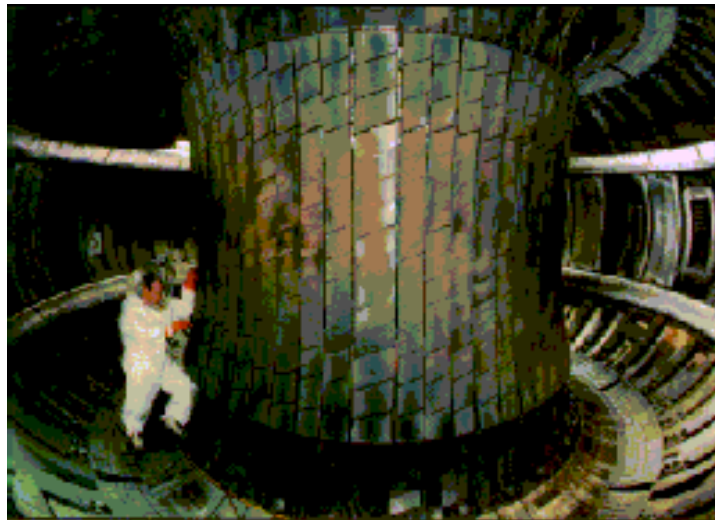
US NARA: ERA

- Currently have 4PB of electronic records.
- Lockheed Martin consortium awarded \$307M contract to develop ERA.
- Delivered system (2007-2011) will:
 - Be an integrated system of COTS products
 - Interface with legacy systems



JET: Processed Pulse File

- 60,000+ pulses, many TB of data
- IBM mainframe with disk farm and tape libraries
- Replaced in 2001:
 - Modular design
 - 2.5M data files converted to open format
 - Commercial RDMS
 - Metadata stored with data to preserve context.



Questions?