

Digital preservation: an introduction

Michael Day
UKOLN, University of Bath, UK
m.day@ukoln.ac.uk

University of the West of England, MSc in Information and Library Management,
Advanced Information Systems module
Frenchay Campus, Bristol, 24th October 2006



<http://www.ukoln.ac.uk/>



Session overview

- Some definitions
- The digital preservation problem
- Preservation strategies
- The OAIS reference model
- Preservation metadata (documentation)
- Non-technical issues
 - Stewardship, collection management, legal issues, costs, ...
- Selected projects and initiatives
- Case study: the World Wide Web



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Definitions



<http://www.ukoln.ac.uk/>



Definitions (1)

- Preservation:
 - A management function
 - “Its objective is to ensure that information survives in usable form for as long as it is wanted” - John Feather (1991)
 - Not *primarily* about:
 - Conservation or restoration
 - Storage media or backup regimes
 - Concepts of “permanence”



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Definitions (2)

- Digital preservation:
 - Digital information is different
 - Technical problems with ensuring continued access (more of this later)
 - But also (primarily) a managerial problem
 - "... the planning, resource allocation, and application of preservation methods and technologies to ensure that digital information of continuing value remains accessible and usable" - Margaret Hedstrom (1998)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Definitions (3)

- Digital curation:
 - New(ish) phrase
 - Concept (data curation) originates in the scientific data world (e.g. bioinformatics, astronomy)
 - Is central to the UK Digital Curation Centre
 - Is used to mean something more than just the preservation of objects
 - "The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and reuse" - Philip Lord, *et al.* (2004)
 - "Maintaining and adding value to a trusted body of information for current and future use" -- DCC presentation at CNI (2005)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Definitions (3)

- Some confusing terminology:
 - “Archiving”
 - A term used in some computing contexts for the creation of secure backup copies
 - Sometimes used (loosely) in preservation contexts
 - “Archives”
 - A well-understood (and discussed) term in archives and recordkeeping professions
 - But is also used informally to refer to almost any collection of digital 'stuff'
 - e.g., e-print archives, image archives, etc.



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Definitions (4)

- Confusing terminology (continued):
 - “Digitisation”
 - The conversion of non-digital objects into digital form
 - The phrase 'digital preservation' was used historically to refer to digitisation where the main motive was the preservation of original items (some care needed when using older literature)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



The digital preservation problem



<http://www.ukoln.ac.uk/>



Storage media (1)

- Media issues:
 - Currently magnetic or optical tape and disks, some devices (e.g., memory sticks)
 - Examples include: CD-ROM, DVD (optical), DAT, DLT (magnetic)
 - Unknown lifetimes
 - Subject to differences in quality or storage conditions
 - But relatively short lifetimes compared to paper or good quality microform
 - Probably years rather than decades



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Storage media (2)

- Media issues (continued):
 - Format differences
 - Technical solutions
 - Longer lasting media:
 - » e.g. Norsam's High Density Rosetta system - analogue storage on nickel plates
 - » COM (output to good-quality microform)
 - » Keeping paper copies!
 - Periodic copying of data bits on to new media (refreshing) - data management solution, e.g. for hierarchical storage systems



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Hardware & software dependence

- Most digital objects are dependent on particular configurations of hardware and software
 - The heart of the digital preservation problem
 - Relatively short obsolescence cycles for:
 - Hardware
 - » e.g., BBC Domesday Project (1986) used a special type of videodisc player developed by Philips
 - Software
 - » e.g., word-processing files
 - Article on the recovery of Domesday Project content:
<http://www.atsf.co.uk/dottext/domesday.html>



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Conceptual problems (1)

- What is an digital object?
 - Some are analogues of traditional objects, e.g. correspondence, research papers
 - Others are not, e.g. Web pages, GIS, 3D models of chemical structures
- Three layers (from: Thibodeau, 2002):
 - Physical: the bits stored on a particular medium
 - Logical: defines how the bits are used by a software application, based on data types (e.g. ASCII); in order to understand (or preserve) the bits, we need to know how to process this
 - Conceptual: things that we deal with in the real world



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Conceptual problems (2)

- On which of these layers should preservation activities focus?
 - We need to preserve the ability to reproduce the objects, not just the bits
 - In fact, we can change the bits and logical representation and still reproduce an *authentic* conceptual object (e.g. converting into PDF)
- Authenticity and integrity
 - How can we trust that an object is what it claims to be?
 - Digital information can easily be changed by accident or design



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Scale (1)

- An increasing flood of data ...
 - The Web
 - Billions of pages
 - Internet Archive - >2 Petabyte (and still growing @ 20 Tb. per month)
 - The "deep-Web"
 - Scientific data
 - Wellcome Trust Sanger Institute - manages several hundred Terabytes of data per year, growing exponentially (just one data centre)
 - Particle physics, Earth Observation and astronomy - e-Science projects expected to generate Petabytes of data per year (e.g., CERN's Large Hadron Collider = ca. >15 Pb)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Scale (2)

- Sizes (broadly):

Kilobyte:	1,000 bytes
Megabyte:	1,000,000 bytes
Gigabyte:	1 billion bytes
Terabyte:	1,000 Gigabytes
Petabyte:	1,000 Terabytes



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Scale (2)

- Sizes (broadly):

Kilobyte:	1,000 bytes
Megabyte:	1,000,000 bytes
Gigabyte:	1 billion bytes
Terabyte:	1,000 Gigabytes
Petabyte:	1,000 Terabytes
Exabyte:	1,000 Petabytes
Zettabyte:	1,000 Exabytes
Yottabyte:	1,000 Zettabytes



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Some general principles (1)

- Most of the technical problems associated with long-term digital preservation can be solved if a life-cycle management approach is adopted
 - i.e. a continual programme of active management
 - Ideally, combines both managerial and technical processes, e.g., as in the OAIS Model
 - Many current systems (e.g. repository software) are attempting to support this approach
 - Preservation strategies need to be seen in this wider context
- Preservation needs to be considered at a very early stage in an object's life-cycle



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Some general principles (2)

- Need to identify and understand the 'significant properties' of an object
 - Focuses on the essential
 - Helps with choosing an acceptable preservation strategy
- Encapsulation may have some benefits
 - Surrounding the digital object - at least conceptually - with all of the information needed to decode and understand it (including software)
 - Produces autonomous 'self-describing' objects, reduces external dependencies; linked to the Information Package concept in the OAIS Reference Model
- Keep the original byte-stream in any case



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Digital preservation strategies



<http://www.ukoln.ac.uk/>



Preservation strategies

- Three main families:
 - Technology preservation
 - Technology emulation
 - Information migration
- Also:
 - Digital archaeology (rescue)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Technology preservation

- The preservation of an information object together with all of the hardware and software needed to interpret it
 - Successfully preserves the look, feel and behaviour of the whole system (at least while the hardware and software still functions)
 - May have a role for historically important hardware
 - Problems with storage and ongoing maintenance, missing documentation
 - Would inevitably lead to 'museums' of "ageing and incompatible computer hardware" -- Mary Feeney
 - May have a short-term role for supporting the rescue of digital objects (digital archaeology)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Technology emulation (1)

- Preserving the original bit-streams and application software; running this on emulator programs that mimic the behaviour of obsolete hardware
- Emulators change over time
 - Chaining, rehosting
 - Emulation Virtual Machines
 - Running emulators on simplified 'virtual machines' that can be run on a range of different platforms
 - Virtual machines are migrated so the original bit-streams do not have to be



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Technology emulation (2)

- Benefits:
 - Technique already widely used, e.g. for emulating different hardware, computer games
 - Preserves the original bits
 - Reduces the need for regular object transformations (but emulators and virtual machines may themselves need to be migrated)
 - Retains 'look-and-feel'
 - May be the only approach possible where objects are complex or dependent on executable code
 - Less 'understanding' of formats is needed; little incremental cost in keeping additional formats



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Technology emulation (3)

– Issues

- Which organisations have the technical skills necessary to implement the strategy?
- Preserving 'look and feel' may not be needed for all objects
- It will be difficult to *know* definitively whether user experience has been accurately preserved

– Conclusions

- Promising family of approaches
- Needs further practical application and research



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Information migration (1)

– Managed transformations

- A set of organised tasks designed to achieve the periodic transfer of digital information from one hardware and software configuration to another, or from one generation of computer technology to a subsequent one - CPA/RLG report (1996)
- Abandons attempts to keep old technology (or substitutes for it) working
- A 'known' solution used by data archives and software vendors (e.g., a linear migration strategy is used by software vendors for some data types, e.g. Microsoft Office files)
- Focuses on the *content* of objects



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Information migration (2)

- Main types (from OAIS Model)
 - Refreshment
 - Replication
 - Repackaging
 - Transformation
- Issues
 - Labour intensive
 - There can be problems with ensuring the 'integrity and authenticity' of objects
 - Transformations need to be documented (part of the preservation metadata)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Information migration (3)

- Uses
 - Seems to be most suitable for dealing with large collections of similar objects
 - Migration can often be combined with some form of standardisation process, e.g., on ingest
 - ASCII
 - Bit-mapped-page images
 - Well-defined XML formats
 - Migration on Request (CAMiLEON project)
 - Keep original bits, migrate the rendering tools



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Digital archaeology

- Not so much a preservation strategy, but the default situation if we fail to adopt one
- Using various techniques to recover digital content from obsolete or damaged physical objects (media, hardware, etc.)
 - A time consuming process, needs specialised equipment and (in most cases) adequate documentation
 - Considered to be expensive (and risky)
 - Remains an option for content deemed to be of value



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Other strategies

- Digital archaeology
 - data recovery
 - time consuming process, needs specialised equipment (i.e., expensive)
- “Persistent archives”
 - San Diego Supercomputer Center
 - Research funded by NSF, DARPA, NARA
 - Comprehensive strategy based on an information management architecture
 - Infrastructure independent representations of digital objects (tagged in XML)
 - Tested on an e-mail collection (Reagan Moore, *et al.*, 2000)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Encapsulation

- Encapsulating the digital object with all of the information needed to decode and understand it
 - Not specific to any particular preservation strategy
 - Self-describing objects
 - The principle underlying the Information Package concept in the OAIS Reference Model (more of this later)
 - Examples:
 - Universal Preservation Format (UPF)
 - “Buckets” (NASA Langley Research Center)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Choosing a strategy (1)

- Preservation strategies are not in competition (different strategies will work together)
 - A suggestion that we should keep the original bits (with documentation) in any case
- But the strategy chosen has implications for:
 - The technical infrastructure required (and metadata)
 - Collection management priorities
 - Rights management
 - e.g. Owning the rights to re-engineer software
 - Costs



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Choosing a strategy (2)

- Decision support tools
 - Preservation strategies
 - Target formats for transformations
- Nationaal Archief (Netherlands) testbed project (general experimental framework)
- Vienna University of Technology tool based on utility analysis (cost-benefit analysis)
- Both developed further by the Digital Preservation cluster of the DELOS Network of Excellence on Digital Libraries
 - <http://www.dpc.delos.info/>



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Case study

Rescue of content from BBC
Domesday videodiscs



<http://www.ukoln.ac.uk/>



Rescue of BBC Domesday (1)

- BBC Domesday project (1986)
 - To commemorate the 900th Anniversary of the original Domesday survey
 - Two interactive videodiscs (12")
 - Mixture of textual material (some produced by schools), maps, statistical data, images and video
 - Technical basis:
 - Hardware: BBC Master Series microcomputer and Philips Laservision (LV-ROM) player
 - Some software in ROM chip, others on the discs
 - System obsolete by end of 1990s; working hardware becoming more difficult to find



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Rescue of BBC Domesday (2)

- CAMiLEON project
 - Proof of concept for the emulation approach
 - Converted data into media-neutral form
 - Adapted an existing emulator for the BBC microcomputer to render Domesday content
- The National Archives (and partners)
 - Reengineered the whole system for use on Windows PCs
 - Digital versions of images and video converted from original master tapes (still held by BBC)
 - Developed an improved interface
 - Web version: <http://domesday1986.com/>



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



http://domesday.domesday1986.com/EXEC - Microsoft Internet Explorer

File Search Help Normal Download / High Quality Images

Domesday 1986 - Community Data

Print Text

LARGE TEXT

TEXT INDEX Southern Britain

01	SOUTH BRITAIN Introduction
02	Classification of Regions
11	Table 1: Growth Areas
12	Table 2: Rural Areas
17	Table 3: Primary Prod. Areas
19	Table 4: Declining Areas
22	Table 5: Resorts
24	Table 6: White-Collar Britain
27	Table 7: Middle Britain
28	The South-East Core Region
33	East Anglia & the South-West
35	The Midlands
39	The North-West & Humberside
42	Wales
43	Conclusion

Darker Lighter

Print Photos

Information about the current map.
Map 18797 Southern Level 1

Information about the current picture.
Picture 18593

Composite Landsat Image

Digital mosaic of Multispectral Scanner Landsat images from 1976-82 in simulated natural colour. Supplied by Nat'l Remote Sensing Cntr, Farnborough



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



http://domesday.domesday1986.com/EXEC - Microsoft Internet Explorer

File Search Help Normal Download / High Quality Images

Domesday 1986 - Community Data

Print Text

LARGE TEXT

0388 0078 TEXT INDEX SY 88 78

01	Kimmeridge and Tynham
02	Contributors

More Photos Darker Lighter

Print Photos

Information about the current map.
Map 19821 Southern Level 3

Grid Refs to one kilometre in Great Britain
Bottom Left 0388 0078 SY 88 78
Top Right 0392 0081 SY 92 81

Information about the current picture.
Picture 0590

Kimmeridge Bay & Tynham Cap

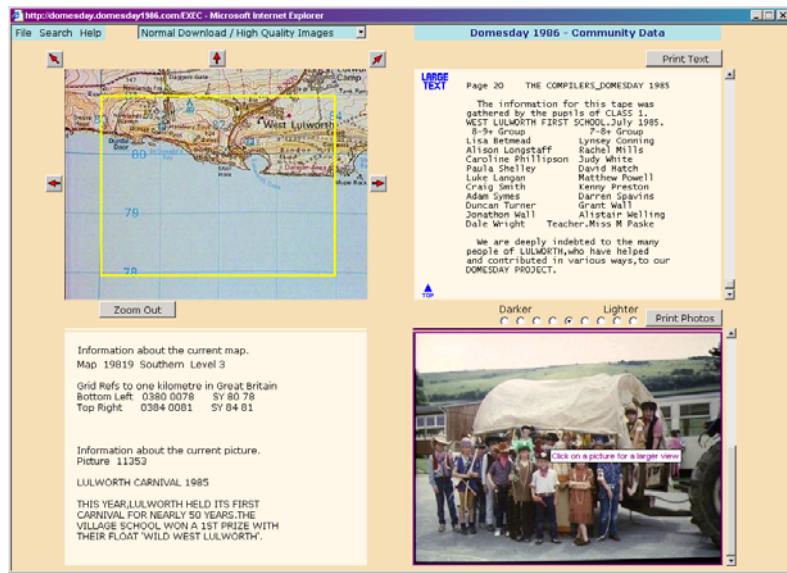
From behind the old coastguard cottages looking WNW across Kimmeridge Bay with Tynham cap in the Background from 03912/00788



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>





Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



The OAIS reference model



<http://www.ukoln.ac.uk/>



The OAIS reference model

- Reference Model for an Open Archival Information System (OAIS)
 - Development managed by the Consultative Committee on Space Data Systems (CCSDS)
 - CCSDS Blue Book 650.0-B-1 (2002)
 - ISO 14721:2003
 - Currently under review
 - Has established a common framework of terms and concepts
 - Information model has been influential on the design of some preservation metadata schemas
 - Conformance ...



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



OAIS mandatory responsibilities

- Negotiating and accepting information
- Obtaining sufficient control of the information to ensure long-term preservation
- Determining the "designated community"
- Ensuring that information is **independently understandable**, i.e. without the assistance of those who produced it
- Following documented policies and procedures
- Making the preserved information available



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



OAIS Functional Model (1)

- Six entities
 - Ingest
 - Archival Storage
 - Data Management
 - Administration
 - Preservation Planning
 - Access
- Described using UML diagrams

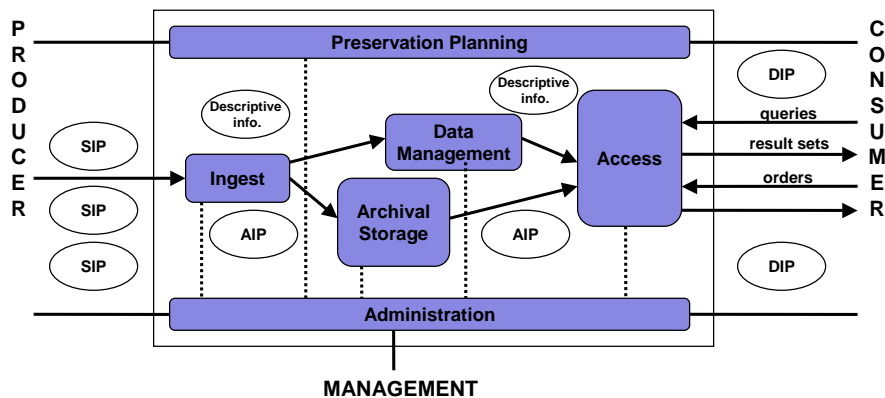


Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



OAIS Functional Model (2)



OAIS Functional Entities (Figure 4-1)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Implementing OAIS

– Fundamentals:

- OAIS is a reference model (conceptual framework), NOT a blueprint for system design
- It informs the design of system architectures, the development of systems and components
- It provides common definitions of terms ... a common language, means of making comparison
- But it does NOT ensure consistency or interoperability between implementations
- Conformance only relates to mandatory responsibilities and following information model



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Preservation metadata



<http://www.ukoln.ac.uk/>



Preservation metadata (1)

- All digital preservation strategies depend - to some extent - on the creation, capture and maintenance of metadata
 - "Preserving the right metadata is key to preserving digital objects" (ERPANET Briefing Paper, 2003)
 - The various types data that will allow the re-creation and interpretation of the structure and content of digital data over time (Ludäscher, Marciano & Moore, 2001)
 - The "information a repository uses to support the digital preservation process," specifically "the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context" (PREMIS Data Dictionary, 2005)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Preservation metadata (2)

- Different roles:
 - "... to find, manage, control, understand or preserve ... information over time" (Cunningham, 2000)
 - Includes most traditional categories of metadata:
 - Descriptive information; technical information about formats and structure; information about provenance and context; administrative information, e.g. for rights management
- There is a perception that different strategies (and objects) will need different metadata



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Preservation metadata - standards

- Existing standards:
 - Either very complex or only provide a basic framework (sometimes both!)
 - Standards developed from many different perspectives:
 - Preservation metadata
 - OCLC/RLG Preservation Metadata Framework, Cedars, NEDLIB, NLA, NLNZ (OAIS influence strongest)
 - METS, NISO Z39.87 (to support digitisation initiatives)
 - PREMIS Data Dictionary (2005)
 - Other relevant standards have also been developed with other aspects of object management in mind:
 - Records management (VERS, RKMS, ISO/DIS 23081-1)
 - Multimedia (MPEG-7, SMPTE)
 - Rights management (MPEG-21)

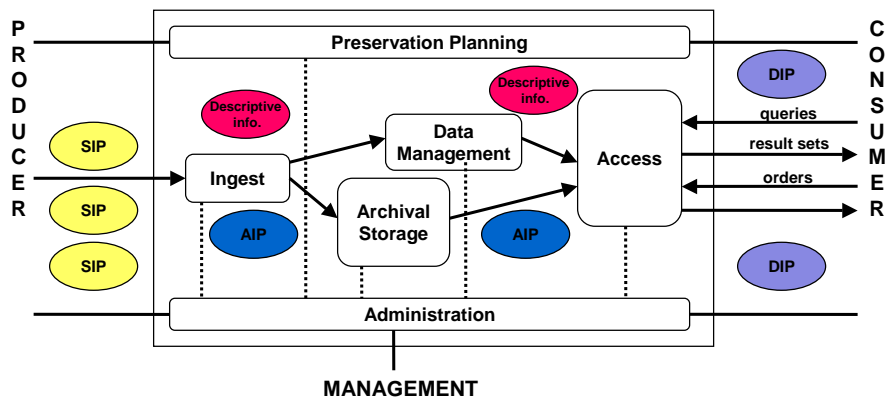


Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



OAIS Functional Model (reprise)



OAIS Functional Entities (Figure 4-1)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



OAIS information objects

- Information Object (basic concept)
 - Data Object (bit-stream)
 - Representation Information (permits “the full interpretation of Data Object into meaningful information”)
- Information Object Classes
 - Content Information
 - Preservation Description Information (PDI)
 - Packaging Information
 - Descriptive Information



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



OAIS information packages

- Information package:
 - Container that encapsulates Content Information and PDI
 - Packages for submission (SIP), archival storage (AIP) and dissemination (DIP)
 - AIP = “... a concise way of referring to a set of information that has, in principle, all of the qualities needed for permanent, or indefinite, Long Term Preservation of a designated Information Object”
 - PDI = other information (metadata) “which will allow the understanding of the Content Information over an indefinite period of time”
 - Reference, Provenance, Context, Fixity



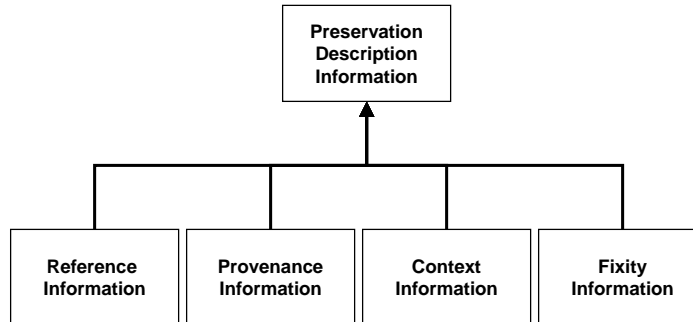
Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



The OAIS model (4)

Preservation Description Information:



OAIS Information Package Taxonomy (Figure 4-14)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



PREMIS Data Dictionary (1)

- Preservation Metadata: Implementation Strategies
- Working Group sponsored by OCLC and RLG
- Reviewed earlier Metadata Framework document and existing practice
- Focus on implementation and definition of 'core' metadata
- PREMIS Data Dictionary (May 2005)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



PREMIS Data Dictionary (2)

- PREMIS Data Dictionary version 1.0
 - Moves away from OAIS Information Model structure
 - Developed own information model
 - Defines semantic units for: Objects, Events (Agents, Rights)
- Also:
 - An XML implementation
- Maintenance activity (led by the Library of Congress)
- PREMIS Implementors Group (PIG)
- Already thinking about v 2.0



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



PREMIS: Preservation Metadata Maintenance Activity (Library of Congress) - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.loc.gov/standards/premis/index.html

The Library of Congress > Standards > PREMIS Home

Standards Pages SEARCH

PREMIS

PRESERVATION METADATA MAINTENANCE ACTIVITY

Official Web Site

- ▶ [Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group](#) [PDF: 3.2MB / 237p.]
- ▶ [PREMIS schemas](#)
- ▶ [Changes to PREMIS data dictionary and schemas](#)
- ▶ [PREMIS Editorial Committee](#) **New!**
- ▶ [PREMIS Implementors' Group \(PIG\)](#)
- ▶ [PREMIS Resources: articles and presentations](#)
- ▶ [PREMIS Implementation Registry](#)
- ▶ [PREMIS Information Sheet](#)
- ▶ [PREMIS Working Group Home Page](#) [OCLC]
- ▶ [PREMIS Implementation Survey](#) [PDF: 1.24MB / 66p.]
- ▶ [Comments](#)

The PREMIS maintenance activity is responsible for maintaining, supporting, and coordinating future revisions to the PREMIS data dictionary. The Preservation Metadata: Implementation Strategies Working Group, convened by OCLC and ELS, initially developed the PREMIS data dictionary as a specification with the goal of creating an implementable set of "core" preservation metadata elements, with broad applicability within the digital preservation community. Supporting XML schemas allow for implementation of the core metadata element set and are maintained in the Network Development and MARC Standards Office of the Library of Congress.

As of May 2005 the PREMIS data dictionary and schemas will begin a period of trial use. It is expected that they will remain stable for at least a year, after which revisions may be made based on results of experimentation.

News and articles:

- ▶ PREMIS wins the 2006 Society of American Archivists' [Preservation Publication Award](#)
- ▶ [Current news](#): PREMIS maintenance activity has commissioned two consultancies
- ▶ [Announcement: PREMIS working group wins 2005 Digital Preservation Award](#)
 - ▶ [Award certificate](#)
- ▶ ["Practical Preservation: the PREMIS Experience"](#)
Priscilla Caplan and Rebecca Guenther
Library Trends: 54 (1) Summer 2005
- ▶ ["Preservation Metadata"](#)
Brian Lavoie and Richard Gartner
DPC Technology Watch Report No. 05-01: September 2005

PREMIS Implementors' Group Forum (pig@loc.gov):

An unmoderated listserv open to members of the PREMIS implementor community. To subscribe to the forum:

1. send email message to:
listserv@loc.gov

Done

Is metadata sustainable?

- Metadata is expensive to create and maintain:
 - There is a need to balance the risks of data loss (or costs of recovery) with the costs of creating metadata
 - Automatic capture of some types of metadata
 - Metadata already embedded in objects or in secondary databases; capture from archive processes
 - Sharing information via registries of format information (Representation Information)
 - Avoid imposing unnecessary costs:
 - Avoid large schemas (?)
 - Need to identify the *right* metadata - 'core metadata' (?)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



The role of registries

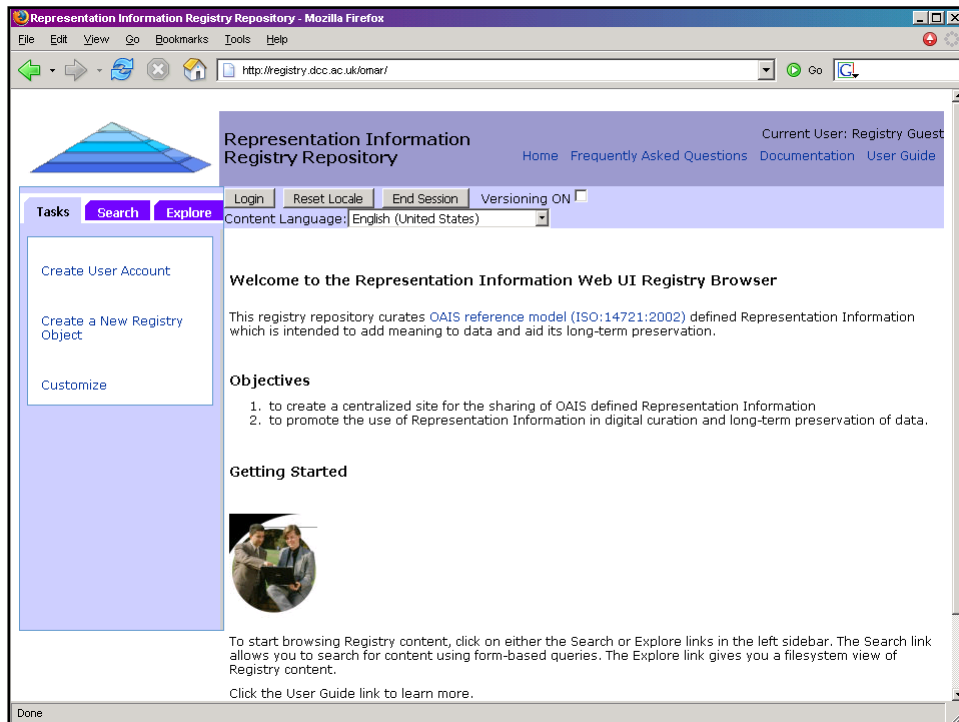
- Registries for sharing information and for identifying or validating formats, etc.
 - There is "... a pressing need to establish reliable, sustained repositories of file format specifications, documentation, and related software" (Lawrence, *et al.*, 2000)
 - DSpace 'bitstream format registry'
 - Digital Library Federation, *et al.* have proposed a Global Digital Format Registry (GDFR)
 - Some components exist, Typed Object Model, JHOVE tool, but GDFR not funded at present
 - Digital Curation Centre Representation Information registry (demonstration)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>





Non-technical issues



<http://www.ukoln.ac.uk/>



Stewardship

- There is widespread confidence that most technical issues can be solved given the organisational will
- Brings us to the most significant set of problems:
 - Stewardship for the long-term needs sustainable institutions willing to take on responsibility for digital content
 - No significant additional funding is likely (but compare NARA's recent multi-million Dollar contract for an Electronic Records Management system)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Collection management

- Selection, storage, access, "de-selection"
- Issues:
 - Preservation issues need to be considered early in an object's life-cycle (the traditional 'transfer to repository at end of active life' model will not work for most objects)
 - An important role for creators (and funding bodies)
 - Guidance, documentation needed
 - Sharing of responsibilities
 - A need for collaboration and infrastructures that support this
 - Digital storage costs are cheap, so should we keep everything?



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Legal issues (1)

- Institutions need to obtain the legal rights to preserve digital objects and make them accessible:
 - e.g., copying, the re-engineering of software
 - identify and negotiate with rights holders?
 - but difficult to identify all rights holders ...
 - safeguard rights
 - part of legal deposit?
 - Monitoring legislation and case law



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Legal issues (2)

- Rights holders want increasing control over content
 - e.g., the extension of copyright periods, licensing of access
 - Digital Millennium Copyright Act (US)
 - European Union Copyright Directive
- Consideration of “dark archives” - repositories without access ...



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Costs

- Still very little known about costs:
 - No widely used economic models
 - No clear idea of who pays?
 - Moore's Law (technology)
 - digital storage densities increase while costs decrease
 - not necessarily applicable to Petabytes of data from e-science projects
 - Identification of cost elements is best approach (for now)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Some projects and initiatives



<http://www.ukoln.ac.uk/>



Digital Curation Centre (1)

- Two main drivers:
 - e-Science, the "data deluge," need for continued access and reuse of data
 - Digital preservation
- Jointly funded by the Joint Information Systems Committee (JISC) and the e-Science Core Programme
 - Outreach, services and development
 - Research programme
- Funding from March 2004 (recently extended)
- Consortium:
 - University of Edinburgh (lead partner), University of Glasgow, Council for the Central Laboratory of the Research Councils, University of Bath (UKOLN)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Digital Curation Centre (2)

- Aims:
 - 'Continuing quality improvement in data curation and digital preservation'
- Main foci:
 - Data as evidential base for science and scholarship
 - Role of data curation & preservation as keys to reproducibility and reuse
 - The worlds of e-learning & scholarly communication



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Digital Curation Centre (3)

- Current activities:
 - Associates Network
 - Linking with existing communities of practice
 - Engaging with active curators
 - Research
 - Edinburgh team - computer science (database) research, e.g. annotation
 - Development
 - Testing of tools, Representation Information registry
 - Services
 - Information events, helpdesk, documentation (briefing papers, curation manual), etc.



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Digital Curation Centre (4)

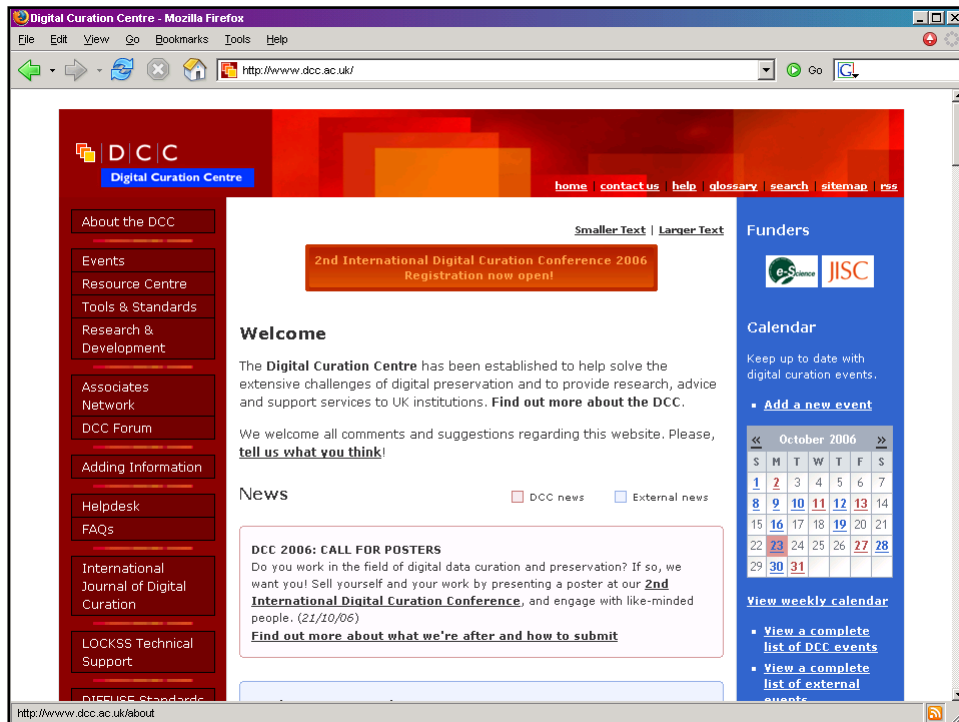
- Progress to date:
 - Web site: <http://www.dcc.ac.uk/>
 - Main source of information about the DCC
 - Various events
 - Thematic workshops
 - 2nd Digital Curation Conference (Glasgow, November 2006)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>





Digital Preservation Coalition

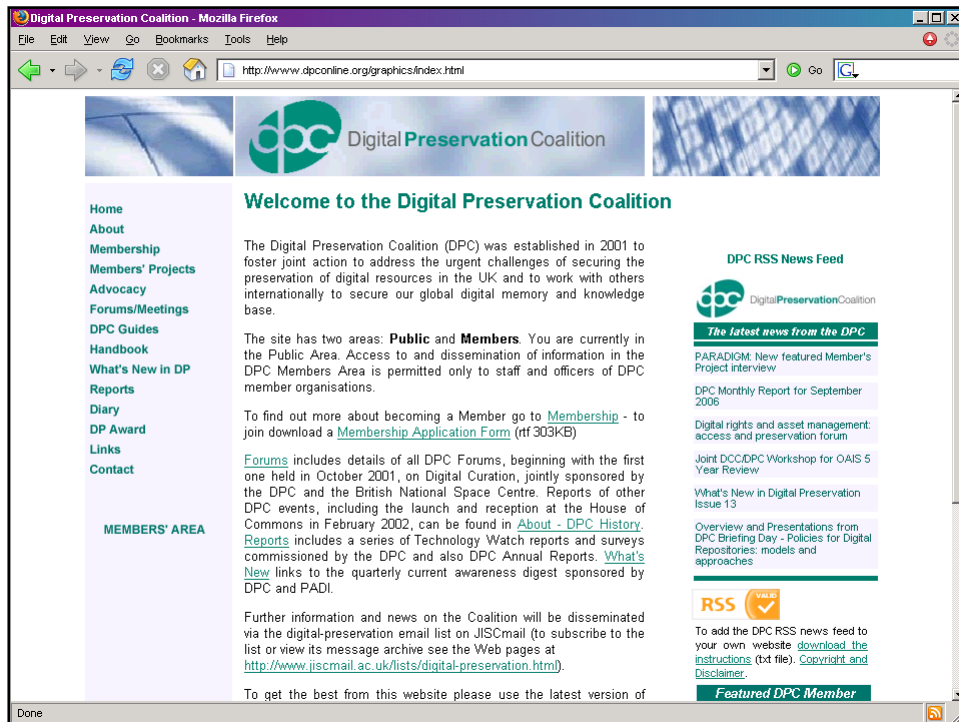
- Formed in 2001
- Aims to foster joint action in the UK and internationally
 - Dissemination
 - Handbook, current awareness bulletin, UK preservation needs assessment (*Mind the Gap!* report)
 - Getting digital preservation on the agenda of key stakeholders
 - Members include BL, JISC, OCLC, The National Archives, MLA, BBC, DCC, etc.
 - Mailing list: digital-preservation@jiscmail.ac.uk
 - <http://www.dpconline.org/>



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>





NDIIPP

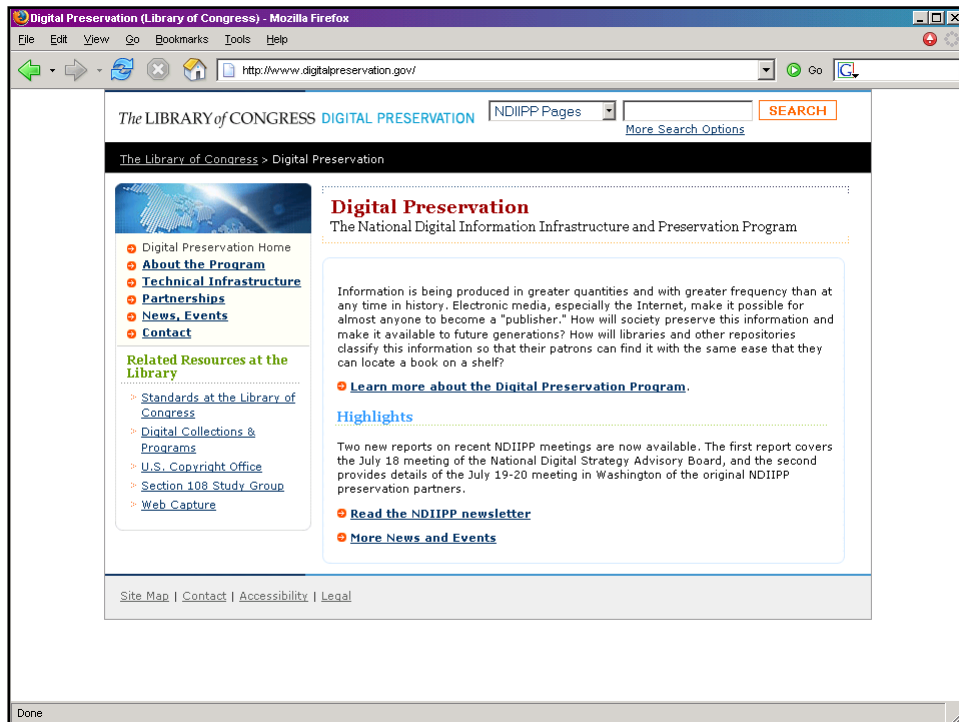
- National Digital Information Infrastructure and Preservation Program
 - Funded by the US Congress
 - A national planning effort led by the Library of Congress, in co-operation with representatives of other federal, research, library, and business organisations
 - Master plan approved by Congress, December 2002
 - Multiple activities:
 - 8 partnership projects (\$14.9 m), from September 2004
 - NSF funded research projects (\$3 m) from 2005 (DIGARCH)
 - <http://www.digitalpreservation.gov/>



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>





Case study

Collecting and preserving Web content



<http://www.ukoln.ac.uk/>



Web archiving - basics (1)

- Web content is transient and ever changing (but is perceived to be of value)
- A very diverse collection:
 - Multiple format types: HTML pages, various image formats, multimedia, etc.
 - Overlaps many traditional 'type' categories:
 - Publications, sound recordings, administrative records, etc.); more informal communication like 'blogs'
 - Problem with sites driven by databases (the deep or hidden Web)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Web archiving - basics (2)

- Problem of 'scale'
 - Multiple Petabytes, repeated
- Significant legal problems:
 - Intellectual property
 - The Web is typically not covered by legal deposit legislation
 - Liability issues with 'republishing' problematic content; e.g. that which is defamatory or otherwise illegal (e.g., some types of pornography)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Web archiving - approaches (1)

- Harvesting
 - Crawling (parts of) the Web with specialised robot programs that download content
 - Works currently for the 'surface Web'
 - Examples:
 - Internet Archive (<http://www.archive.org/>)
 - European Archive
 - National libraries (collecting 'national' Web domains - Swedish Royal Library (Kulturarw³), + other initiatives in Austria, Denmark, Finland, France, Iceland, Norway, Slovenia, etc.
 - National archives initiatives (UK, USA, Australia)
 - International Internet Preservation Consortium (IIPC)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Web archiving - approaches (2)

- Selective approach
 - Web sites of value selected, rights negotiated for collection, sites collected by harvesting (e.g. using capture tools like HTTrack or Heritrix) or by deposit
 - Does not scale to whole Web
 - Examples:
 - National Library of Australia (PANDORA)
 - US National Archives and Records Administration (deposit based)
 - The National Archives (joint project with Internet Archive)
 - UK Web Archive Consortium (<http://www.webarchive.org.uk/>)
 - British Library, National Library of Scotland, National Library of Wales, The National Archives, Wellcome Library, JISC



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Web archiving - approaches (3)

- Selective approach (continued)
 - A major focus on 'events,' e.g.:
 - PANDORA (Sydney Olympic Games)
 - Internet Archive Special Collections (US Presidential elections, 9/11 collection, Hurricane Katrina)
 - National Archives and Records Administration (snapshots of US federal agencies and departments at the end of 2001 - the end of the Clinton era)
 - The National Archives (test capture of No. 10, Downing Street site (2001); UK Central Government Web Archive - co-operation with the Internet Archive for the periodic harvesting of selected central government Web sites)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Web archiving - approaches (4)

- Combined approaches
 - Crawler-based harvesting of surface Web combined with the 'deposit' of deep Web content
 - Examples:
 - Bibliothèque nationale de France
 - Beginning to be followed by other national libraries, sometimes contracting the actual harvesting (and storage) to the Internet Archive (e.g. NLA)
 - Being considered by the British Library in addition to (selective) UK Web Archiving Consortium work - assuming that deposit legislation is in place



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Web archiving - approaches (5)

- International Internet Preservation Consortium
 - A focus of co-operation between the Internet Archive and national and research libraries
 - Development of standards and tools
 - Mostly dealing with the problem of scale
 - Heritrix crawler
 - A standardised storage format (WARC)
 - A user interface (WERA) and search facility (NutchWAX)
 - Standard metadata - automatically documenting selection criteria, the context of retrieval
 - <http://netpreserve.com/>



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Web archiving - some trends

- Most existing initiatives are more concerned with collecting content than with either access or preservation
 - But access issues have been considered by:
 - Internet Archive (e.g., Wayback Machine)
 - Nordic Web Archive project
 - International Web Archiving Consortium
 - PANDORA Archive (NLA)
 - UK Web Archiving Consortium
 - Most captured content can be viewed in the latest generation of Web browsers
 - For example, Wayback Machine: <http://www.archive.org/>



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Internet Archive - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.archive.org/index.php

ARCHIVE Web | Moving Images | Texts | Audio | Software | Education | Patron Info | About Us
Forums | FAQs | Contributions | Jobs | Donate

Search: All Media Types

Universal access to human knowledge

Anonymous User [login](#) or [join us](#)


Announcements [\(more\)](#)

[New Collections this Month](#)

[Bookmark Explorer](#)

[Datacenter moved and settled](#)

Web 55 billion pages

 [Advanced Search](#)

Welcome to the Archive [RSS](#)

The Internet Archive is building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public.


Moving Images 42,888 movies

[Browse](#) [\(by keyword\)](#)
[Upload your own movie](#)

This Just In [\(more\)](#) [RSS](#)

[Midnight Manhunt](#)
41 minutes ago

Curator's Choice [\(more\)](#)

 [The History of Marijuana - Woody Harrelson](#)
How the government and FDA


Live Music Archive 39,367 concerts

[Browse](#) [\(by band\)](#)
[Upload your own concert](#)

This Just In [\(more\)](#) [RSS](#)

[Blues Traveler Live at...](#)
36 minutes ago

Curator's Choice [\(more\)](#)

 [Mike Doughty Live at Plush on 2005-09-27](#)
Disc One 01. intro 02. Tremendous Brunettes 03.


Audio 101,800 recordings

[Browse](#) [\(by keyword\)](#)
[Upload your own recording](#)

This Just In [\(more\)](#) [RSS](#)

[Foro RRHH](#)
50 minutes ago

Curator's Choice [\(more\)](#)

 [Oeuf Korreckt - Scrambled Oeuf \[n011\]](#)
This classic No Type EP revisits


Texts 33,971 texts

[Browse](#) [\(by keyword\)](#)
[Upload your own text](#)

This Just In [\(more\)](#) [RSS](#)

[voz](#)
1 hour ago

Curator's Choice [\(more\)](#)


 [Dictionnaire raisonné de l'architecture...](#)

Done

Internet Archive Wayback Machine - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://web.archive.org/web/*http://www.uwe.ac.uk



Enter Web Address: All [Adv. Search](#) [Compare Archive Pages](#)

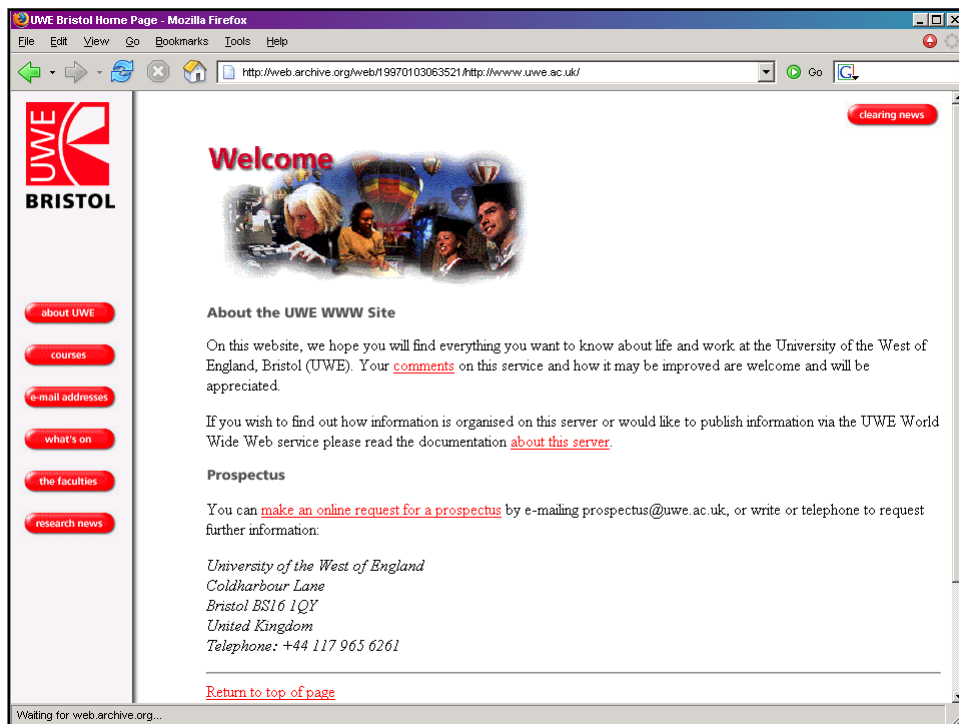
Searched for [http://www.uwe.ac.uk](#) 172 Results

Note some duplicates are not shown. [See all](#).
* denotes when site was updated.

Search Results for Jan 01, 1996 - Oct 23, 2006

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
0 pages	7 pages	2 pages	5 pages	15 pages	0 pages	1 pages	20 pages	70 pages	32 pages	0 pages
Jan 03, 1997 Jan 31, 1997 Jan 31, 1997 Mar 15, 1997 Apr 27, 1997 Apr 28, 1997 Dec 10, 1997	Feb 24, 1998 May 29, 1998	Jan 17, 1999 Jan 25, 1999 Feb 08, 1999 Apr 21, 1999 Apr 27, 1999	Mar 01, 2000 Mar 02, 2000 Apr 07, 2000 Apr 14, 2000 Apr 18, 2000 Apr 20, 2000 May 02, 2000 May 10, 2000 May 11, 2000 May 19, 2000 Jun 19, 2000 Jun 21, 2000 Aug 16, 2000 Aug 24, 2000 Oct 17, 2000			Dec 21, 2002	Jan 25, 2003 Jan 27, 2003 Feb 02, 2003 Apr 27, 2003 Mar 21, 2003 Jun 02, 2003 Apr 02, 2003 Jun 03, 2003 Apr 24, 2003 May 27, 2003 Jun 10, 2003 Jun 12, 2003 Jun 21, 2003 Jul 24, 2003 Aug 01, 2003 Sep 25, 2003 Oct 07, 2003 Oct 08, 2003 Oct 28, 2003 Dec 03, 2003 Dec 04, 2003 Dec 15, 2003 Dec 27, 2003	Feb 05, 2004 Mar 22, 2004 Apr 27, 2004 Jun 02, 2004 Jun 03, 2004 Jun 04, 2004 Jun 08, 2004 Jun 10, 2004 Jun 12, 2004 Jun 14, 2004 Jun 15, 2004 Jun 16, 2004 Jun 18, 2004 Jun 19, 2004 Jun 23, 2004 Jun 24, 2004 Jun 25, 2004 Jun 26, 2004 Jun 27, 2004 Jun 28, 2004 Jun 29, 2004	Feb 03, 2005 Feb 04, 2005 Feb 05, 2005 Feb 06, 2005 Feb 06, 2005 Feb 07, 2005 Feb 11, 2005 Feb 12, 2005 Feb 13, 2005 Feb 17, 2005 Feb 18, 2005 Feb 22, 2005 Feb 23, 2005 Feb 26, 2005 Feb 28, 2005 Mar 03, 2005 Mar 04, 2005 Mar 06, 2005 Mar 07, 2005 Mar 08, 2005 Mar 09, 2005	

Done



Compare with ...



<http://www.ukoln.ac.uk/>





Summing up



<http://www.ukoln.ac.uk/>



Summing up:

- Digital preservation is an organisational as well as a technical problem
- Progress has been made on addressing the technical problems
 - e.g., sustainable preservation strategies and preservation metadata schemas
- However, many other problems remain
- In the longer-term, international co-operation will be essential
 - Some progress made on the national level, e.g. the DPC, DCC, NDIIPP (USA), nestor (Germany)



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Further information



<http://www.ukoln.ac.uk/>



Readings (1)

Neil Beagrie and Maggie Jones, *Preservation Management of Digital Materials: a Handbook* (2001). Updated version available at:
<http://www.dpconline.org/>

Council on Library and Information Resources, *Building a National Strategy for Preservation: Issues in Digital Media Archiving* (April 2002) <http://www.clir.org/pubs/abstract/pub106abst.html>

Council on Library and Information Resources, *The state of digital preservation: an international perspective* (July 2002)
<http://www.clir.org/pubs/abstract/pub107abst.html>

Margaret Hedstrom, *It's about time: research challenges in digital archiving and long-term preservation* (2003)
<http://www.digitalpreservation.gov/>

Margaret Hedstrom and Seamus Ross, *Invest to save: report and recommendations of the NSF-DELOS Working Group on Digital Archiving and Preservation* (2003)
<http://eprints.erpanet.org/archive/00000095/>



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



Readings (2)

Philip Lord and Alison Macdonald, *Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision* (2003) <http://www.jisc.ac.uk/>

Helen R. Tibbo, "On the nature and importance of archiving in the digital age." *Advances in Computers* 57 (2003): 1-67.

Brian Lavoie and Lorcan Dempsey, "Thirteen Ways of Looking at ... Digital Preservation." *D-Lib Magazine* 10, no. 7/8 (July/August 2004)
<http://www.dlib.org/dlib/july04/lavoie/07lavoie.html>

National Science Board, *Long-lived digital data collections: enabling research and education in the 21st century* (2005)
<http://www.nsf.gov/pubs/2005/nsb0540/>

DCC Digital Curation Manual (2005-)
<http://www.dcc.ac.uk/resource/curation-manual/chapters/>



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



More information

Preserving Access to Digital Information
(PADI) gateway:

<http://www.nla.gov.au/padi/>

DPC/PADI "What's New" bulletin:

<http://www.dpconline.org/graphics/whatsnew/>

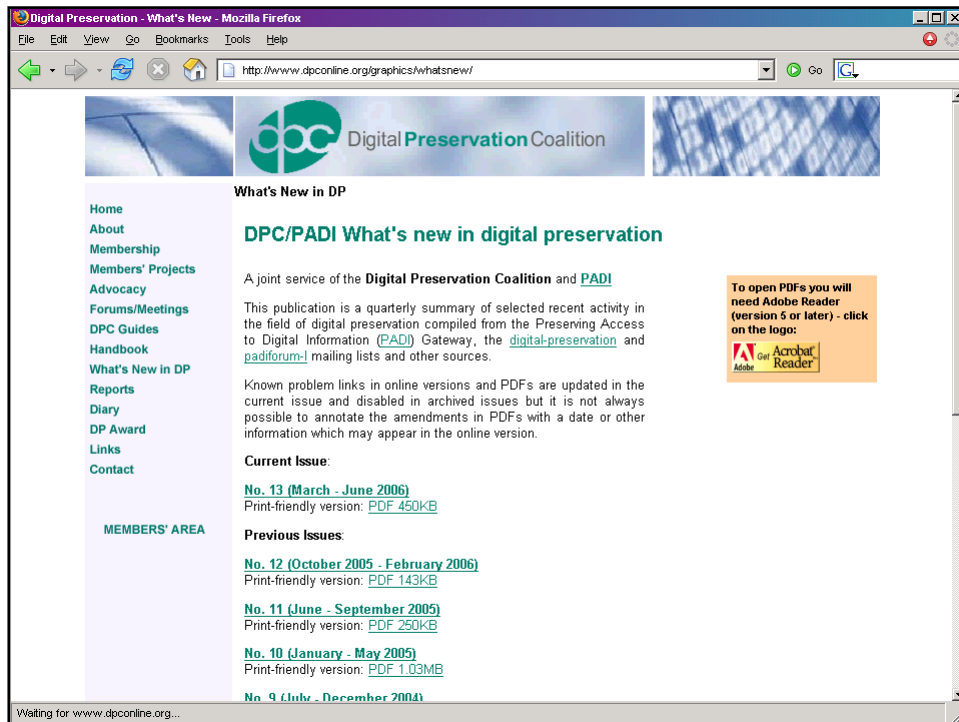


Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



A screenshot of the PADI (Preserving Access to Digital Information) website as it appeared in a Mozilla Firefox browser window. The browser's address bar shows the URL "http://www.nla.gov.au/padi/". The website header features the "PADI" logo in large blue letters, with the tagline "Preserving Access to Digital Information" underneath. To the right of the logo, it says "NATIONAL LIBRARY OF AUSTRALIA". Below the header, a blue sidebar on the left contains a list of navigation links: Home, About PADI, Search, Browse Topics, Feedback, Contributions, What's New?, Partners, and Working Groups. The main content area has a blue background and contains the text "PADI is a subject gateway to international digital preservation resources". Below this text, there are two columns of links. The left column is titled "RESOURCE TYPES" and lists: Events, Policies, Strategies & Guidelines, Projects, Organisations & Websites, Bibliographies, Discussion Lists, Glossaries, Journals & Newsletters, and News & Discussion. The right column is titled "DIGITAL PRESERVATION TOPICS" and lists: Data Documentation & Standards, Digital Libraries, Digital Records, Digitisation, Formats & Media, General Resources, Issues, Management, National Approaches, Rights Management, and Strategies. At the bottom of the left column, there is a link to "padiforum-l". Above the "RESOURCE TYPES" and "DIGITAL PRESERVATION TOPICS" sections, there are two buttons: "PADI TRAILS" with a sub-link "Don't know where to start? Try the new PADI Trails." and "Historical" with a sub-link "New PADI feature!".



Acknowledgements

UKOLN is funded by the Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the UK higher and further education funding councils, as well as by project funding from the JISC, the European Union and other sources. UKOLN also receives support from the University of Bath, where it is based: <http://www.ukoln.ac.uk/>



The Digital Curation Centre is funded by the JISC and the UK e-Science Programme: <http://www.dcc.ac.uk/>



Advanced Information Systems, 24 October 2006

<http://www.ukoln.ac.uk/>



