The digital preservation technological context

Michael Day,
Digital Curation Centre
UKOLN, University of Bath
m.day@ukoln.ac.uk

Preservation of Digital Heritage: Basic Concepts and Main Initiatives, Madrid, 14-16 March 2006

Session overview

- Introductory comments
- Technical issues
- Preservation strategies
- Preservation metadata and shared infrastructure

Introductory comments

Digital preservation (1)

- Concerns continued access (and use)
- Digital preservation is NOT just about technology
- Unites a range of interrelated issues:
  - "... 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)

Digital preservation (2)

- Is sometimes now characterised as 'digital stewardship' or 'digital curation'
  - The concept of data curation originated in data-rich scientific domains like bioinformatics
  - Curation - "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)

The fragility of digital content

The main technical issues

General comments

- Digital information is dependent on its technical environment
- Physical objects are subject to:
  - Physical deterioration
  - Technology obsolescence
- Relatively short timescales

Storage media (1)

- A major focus of concern in the 1970s and 1980s
- Current media types
  - Typically, magnetic or optical tape and disks, various 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

Storage media (2)

- Technical solutions:
  - Periodic copying of data bits on to new media or types of media (refreshing)
  - Longer lasting media
  - Migrating to good-quality microform or paper (!)
- In an organised preservation system, regular routines (quality checking, backup, replication, refreshing, etc.) will help solve the media longevity issue

Technology obsolescence (1)

- A set of much bigger problems
- Software dependence
  - Digital content is, at least in part, dependent on the configurations of hardware and software (applications and operating systems) that were originally used to interpret or display them
- Hardware and software obsolescence
  - Application software and operating systems are upgraded regularly
  - Hardware becomes obsolete or needs repair

Technology obsolescence (2)

- Technical solutions
  - Various preservation strategies have been developed to cope with the obsolescence problem
  - For the most part, these depend on the existence of a continual programme of active management (life cycle management)
  - Supported by systems that implement the various functional entities identified by the Reference Model for an Open Archival Information System (OAIS)
  - Preservation strategies can only be seen in this wider context

Layers of meaning (1)

- Digital objects are logical entities not fixed to any one particular physical carrier
- Three layers (Thibodeau, 2002):
  - Physical objects: the actual bits stored on a particular medium
  - Logical objects: defines how these bits are used by application software, based on data types (e.g. ASCII); in order to understand (or preserve) the byte-streams, we need to know how to process them
  - Conceptual objects: what humans deal with in the real world, meaningful units of information

Layers of meaning (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 could change the bits and logical representation and still reproduce an authentic conceptual object

Authenticity and integrity

- Digital information can easily be changed (e.g., by design or accident)
- How can we trust that an object is what it claims to be?
- Mechanisms are available at the bit level (e.g. checksums), but will this be sufficient?

Problems of scale

- An increasing flood of 'born-digital' data
  - Data deluge in science and engineering
    - » Petabytes generated by high throughput instruments, streamed from sensors and satellites, etc.
  - The World Wide Web
    - » Comprises billions of pages + "deep Web"
    - » Internet Archive = >1 petabyte, and growing @ 20 Tb. per month (http://www.archive.org/)
  - 5 exabytes of *new* information created in 2002:
    - » http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

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

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

Digital preservation strategies

Preservation strategies

- Three main families:
  - Technology preservation
  - Technology emulation
  - Information migration

- Also:
  - Digital archaeology (rescue)

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)

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

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

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

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

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)

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

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

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

Choosing a strategy (2)

- Tools for supporting preservation decisions, e.g.
  - Preservation strategies
  - Target formats for transformations
- Nationaal Archief (Netherlands) testbed project
- Vienna University of Technology utility analysis tool
- Both developed further by the Digital Preservation cluster of the DELOS Network of Excellence

Case study

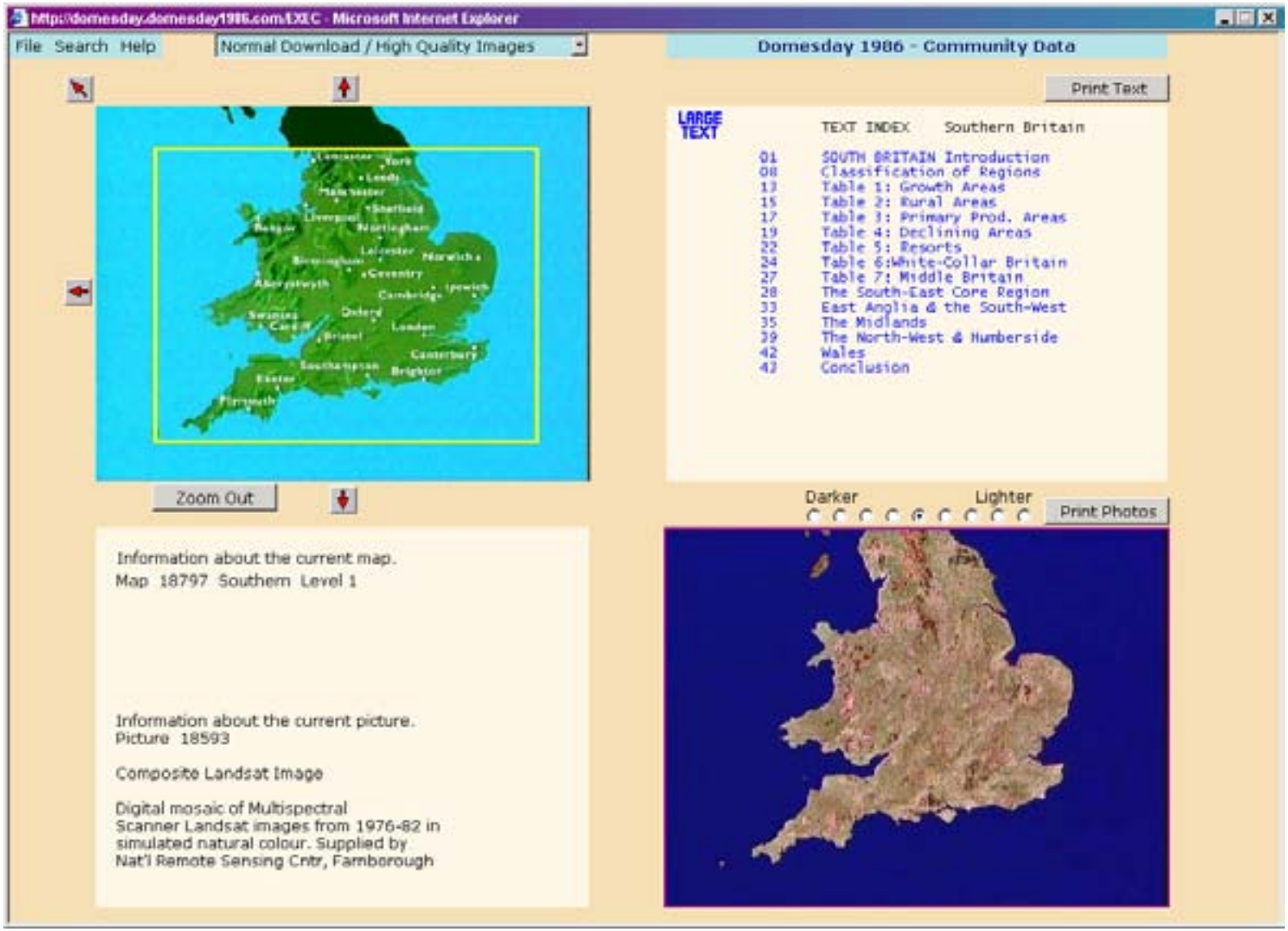
Rescue of content from BBC Domesday videodiscs

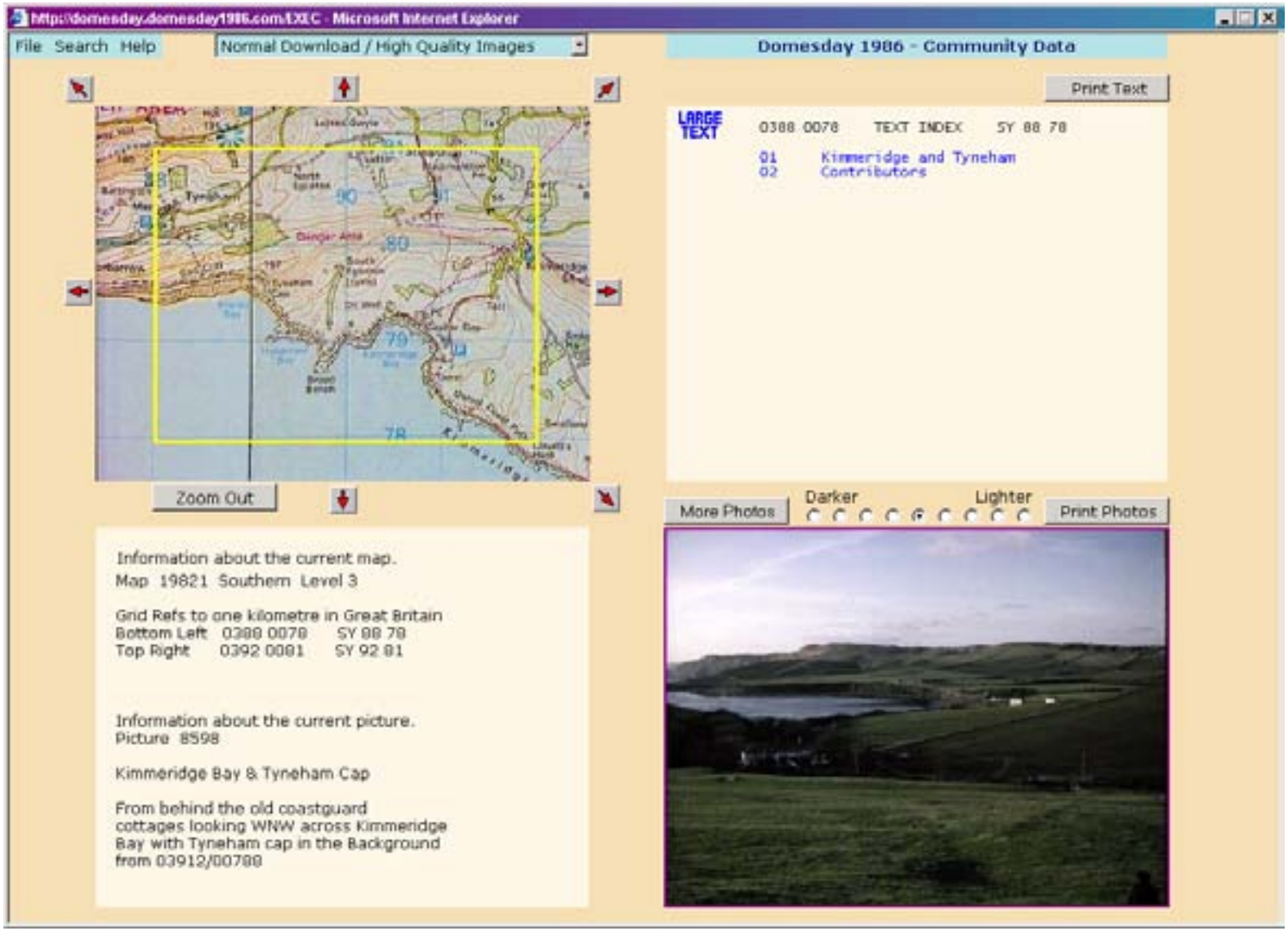
Rescue of BBC Domesday (1)

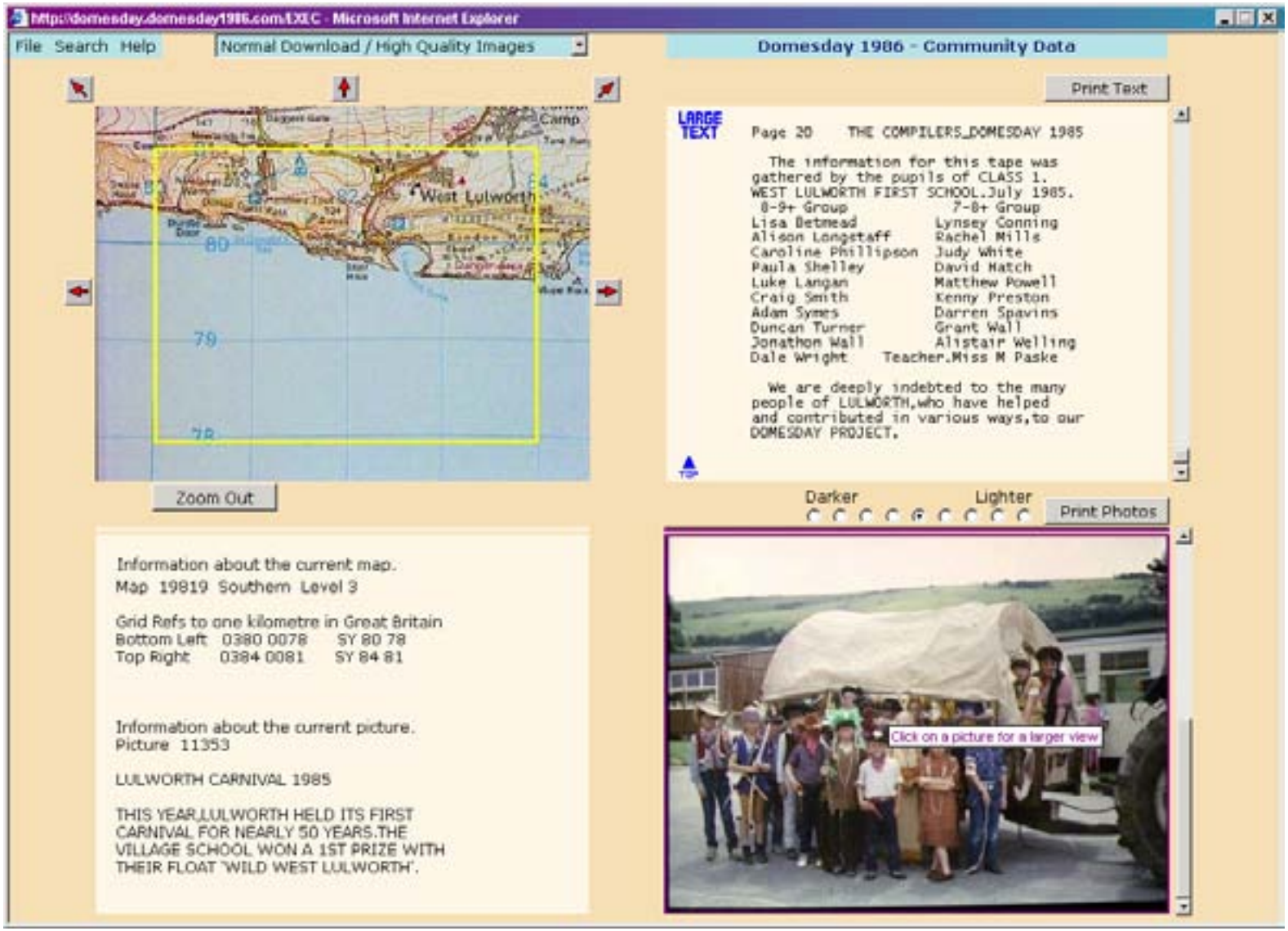
- BBC Domesday project (1986)
  - To commemorate the 900th Anniversary of the original Domesday survey
  - Two interactive videodiscs (12")
    - Mixture of textual material (some produced by schools), maps, statistical data, images and video
  - Technical basis:
    - Hardware: BBC Master Series microcomputer 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

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/


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
08 Classification of Regions
13 Table 1: Growth Areas
15 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
Zoom Out
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


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 Tyneham
02 Contributors
Zoom Out
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 8598
Kimmeridge Bay & Tyneham Cap
From behind the old coastguard cottages looking WNW across Kimmeridge Bay with Tyneham cap in the Background from 03912/00788


http://domesday.domesday1986.com/EXEC - Microsoft Internet Explorer
File Search Help
Normal Download / High Quality Images
Domesday 1986 - Community Data
Print Text
LARGE TEXT
Page 20 THE COMPILERS_DOMESDAY 1985
The information for this tape was gathered by the pupils of CLASS 1. WEST LULWORTH FIRST SCHOOL.July 1985.
8-9+ Group 7-8+ Group
Lisa Betmead Lynsey Conning
Alison Longstaff Rachel Mills
Caroline Phillipson Judy White
Paula Shelley David Match
Luke Langan Matthew Powell
Craig Smith Kenny Preston
Adam Symes Darren Spavins
Duncan Turner Grant Wall
Jonathon Wall Alistair Welling
Dale Wright Teacher.Miss M Paske
We are deeply indebted to the many people of LULWORTH,who have helped and contributed in various ways,to our DOMESDAY PROJECT.
West Lulworth
Zoom Out
Darker
Lighter
Print Photos
Information about the current map.
Map 19819 Southern Level 3
Grid Refs to one kilometre in Great Britain
Bottom Left 0380 0078 SY 80 78
Top Right 0384 0081 SY 84 81
Information about the current picture.
Picture 11353
LULWORTH CARNIVAL 1985
THIS YEAR,LULWORTH HELD ITS FIRST CARNIVAL FOR NEARLY 50 YEARS.THE VILLAGE SCHOOL WON A 1ST PRIZE WITH THEIR FLOAT 'WILD WEST LULWORTH'.
Click on a picture for a larger view

Preservation metadata and shared infrastructures

Preservation metadata (1)

- All digital preservation strategies depend - to a greater or lesser extent - on the creation, capture and maintenance of metadata
  - Preservation metadata:
    - 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)
    - Cuts across older categorisations of metadata (descriptive, administrative, structural)

Preservation metadata (2)

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

Preservation metadata (3)

- PREMIS Data Dictionary
  - Less explicitly based on OAIS Information Model structure than older OCLC/RLG Framework
  - Based on own data model
  - Defines some of the semantic units for: Objects, Events, Agents, Rights
  - Supports automatic capture, where possible
- PREMIS also provides:
  - An XML implementation, e.g. for use in a packaging format like METS (Metadata Encoding and Transmission Standard)
  - Maintenance activity (Library of Congress)

Shared infrastructures

- For example: registries for sharing information about, or 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'
  - Global Digital Format Registry (GDFR)
    - » Some components exist, e.g. Typed Object Model, JHOVE tool
  - DCC Representation Information registry

Some final comments

- The technical issues of digital preservation are only one part of a multidimensional problem
- Progress has been made on addressing technical problems
- Need for sustainability and co-operation
- Need for people with the appropriate skills

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/