IBM Haifa Research Lab

Long Term Digital Preservation

November 2008

Michael Factor

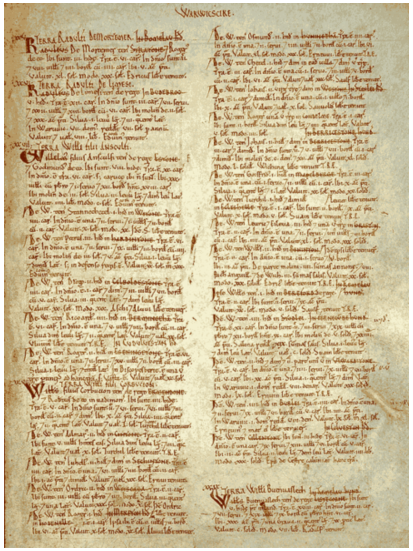
contact: factor@il.ibm.com

http://www.haifa.il.ibm.com/projects/storage/ltdp/index.shtml

The Domesday Book: 1086 vs. 1986

- The original Domesday Book
  - A survey of England completed in 1086 under order of William the Conqueror
  - It is still preserved today since it used (de facto) standards
    - Bit Preservation:
      - Physical Media: Royal Treasury, Chapter House, rarely used
      - Physical Format: Bound Parchment
    - Logical Preservation
      - Logical Format (Representation): Latin (with some vernacular)
      - Context: Anglo-Saxon Chronicle
      - Provenance: Physically tracked

In 2002, after 16 years, the BBC Domesday Project was almost obsolete



http://en.wikipedia.org/wiki/Image:Domesday_book_e31-2-2-f243.gif

- BBC Domesday Project
  - A multi-media edition of the Domesday book published on the 900th anniversary
    - High involvement of schools
  - It was developed using proprietary technologies
    - Bit Preservation:
      - Used an adapted laserdiscs in the LaserVision Read Only Memory
      - Disks viewed on BBC Master microcomputer
    - Information Preservation
      - Software written in BCPL
      - Images stored as single frame analogue overlaid by computer

- Standards and BBC Domesday Project – What Went Wrong?
  - LaserVision ROM vs CD-ROMs
    - At the time CD-ROMs had limited capacity and no standards
  - BBC Microcomputer vs. IBM Compatible PC
    - IBM Compatible's were only used in business at the time.
  - Software written in BCPL
    - Time of language wars – BCPL was a precursor of C
  - Images stored as single frame analogue overlaid by computer
    - JPEG standard from 1992
- The BBC Domesday Project was ahead of its time

It exists today only due to heroic efforts from the CAMiLEON project

http://en.wikipedia.org/wiki/Image:NavComm1.jpg

What is Long Term Digital Preservation?

- *Long Term Digital Preservation (LTDP)* is a means of keeping digital information such that the same information can be used at some point in the future in spite of obsolescence of everything: hardware, software, processes, format, people, etc.
- *Bit Preservation* addresses obsolescence of hardware
  - As the term is used, *digital archiving,* at best, provides bit preservation and makes implicit assumptions on the availability of compatible software, formats, processes
- *Information or logical preservation* addresses obsolescence of everything else

Physical vs. Digital Preservation

	Physical	Digital
Lifetime of the medium	> Centuries	< Decades
Lifetime of the physical form factor	> Millennium	< Decades
Ability to extract the object from the medium	> Millennium	< Decades
Ability to read the object	> Centuries	< Decades
Understanding the object's context	< Decades	< Decades
Knowing the object's provenance	< Decades	< Decades
Ensuring the integrity of the object	Hard	Very Hard
Preserving the preservation system	Not Relevant	Very Hard

- Printing to ensure future usability is not an option
- Accidental physical preservation is possible
  - Accidental digital preservation is not possible

Is Long Term Digital Preservation Needed?

Finance

Rule 17a-4 requires broker-dealers to retain account record information for six years. The six-year period begins either at the time the account is closed or when the information is replaced or updated

Life insurance policies have to be kept for life of policy plus 6-10 years

Healthcare

X-rays are often stored for periods of 75 years

The retention requirement for the [medical] records of minors is 20 to 43 years of age

OSHA requires employers to keep records of . . . employees who are exposed to toxic substances and harmful agents for 30 years

M&E

Film Masters, Out takes. Related artifacts (e.g., games). 100 Years or more

Petroleum

Oil-field data is used over life of field (50+ years)

Pharma

Pharma needs off-line electronic data storage for 50 to 100 years or longer

Aerospace

Aircraft designs records have to be retained for the lifetime of aircraft (50+ years)

Government

Land registry records, social security records, etc. Life of individual to forever

Scientific and Cultural

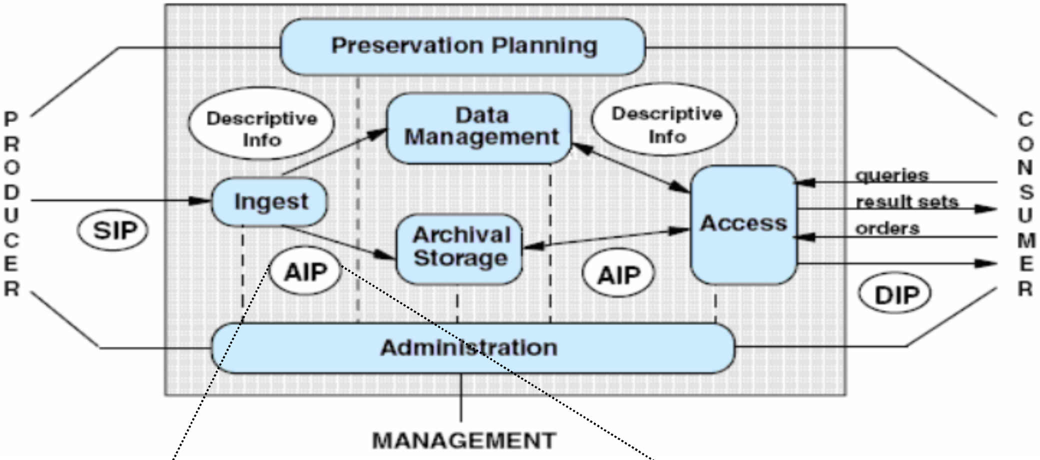
Satellite data is kept for ever

We would like to keep Libraries and Art data for ever

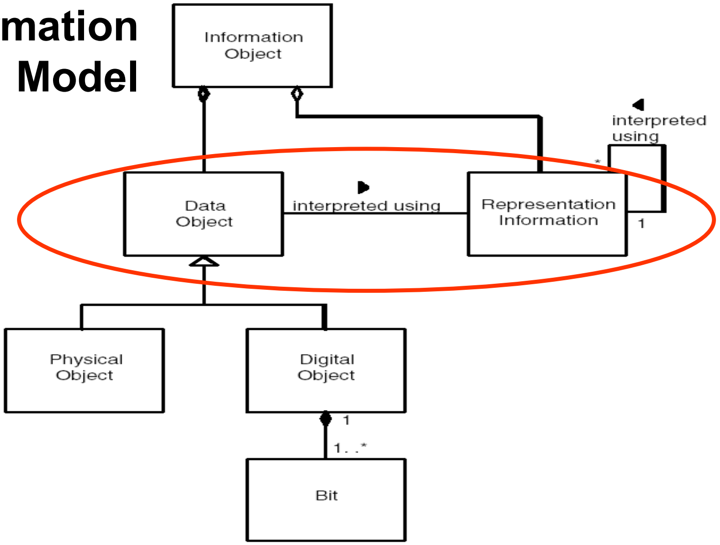
Open Archival Information System (OAIS)

- ISO standard reference model (ISO:14721:2002)
- Provide fundamental ideas, concepts and a reference model for long-term archives
- Incorporate emulation, migration, descriptive via encapsulation
- Focused on logical preservation

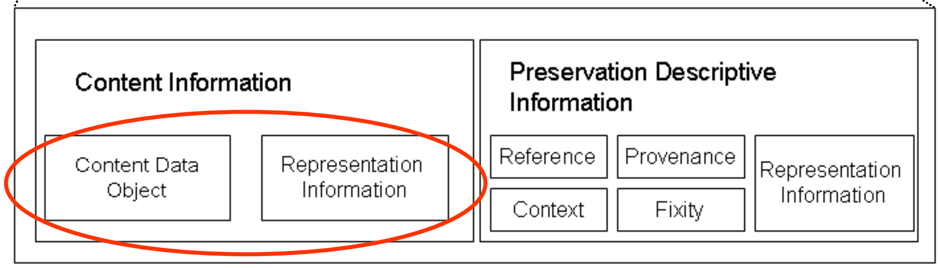
Functional Model



Information Model



AIP

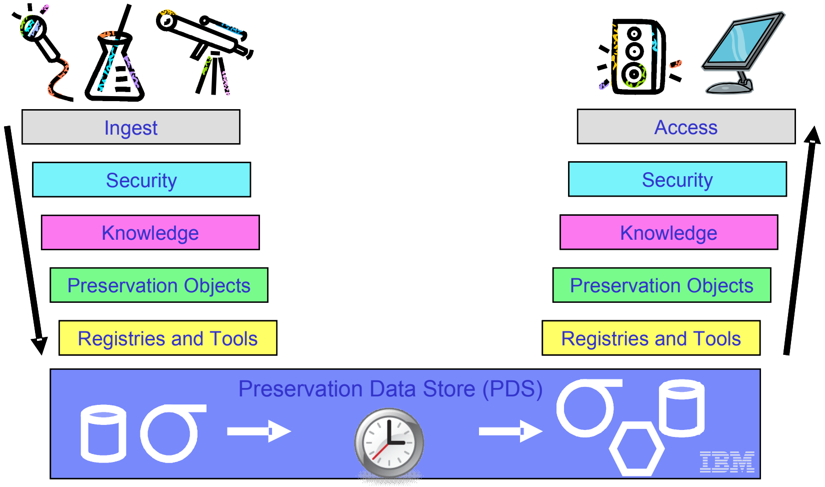

Content Information
Content Data Object
Representation Information
Preservation Descriptive Information
Reference
Provenance
Context
Fixity
Representation Information

Preservation Approaches

Approach	Description	Pros	Cons
Museum	Content and rendering devices are preserved in their original state and maintained operational	No loss of information	Expensive; time bounded; not scalable
Emulation	Adapt the rendering device by emulating it to up-to date software and computers	Problem reduced to preserving the emulation platform; cost proportional to number of formats.	Upfront investment; Only for data coupled with software; Does not allow new interpretations.
Migration	Migrate to newer formats	Less investment when data ingested. Allows new uses.	May introduce noise; cost proportional to data size; continuous cost
Descriptive	Add metadata to fully describe representation of data, allowing writing code in the future to process format	No loss of info; Minimal assumptions on future. Delay's cost until needed	Doesn't support proprietary formats. May have future high cost
Encapsulation	Group together the data and related metadata (including instructions to enable future interpretation)	Most flexible; consistent with everything but museum approach; OAIS compliant	Doesn't tell you what to do

CASPAR and Preservation DataStores

- CASPAR: Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval
  - 8.8M Euro, 3.5 year, EU Projec
  - IBM is the largest IT partner
  - http://www.casparpreserves.eu/
- Demonstrate validity of OAIS framework with heterogeneous data
- IBM is responsible for the storage infrastructure
  - Developed Preservation DataStores
- IBM's Experience in CASPAR
  - Learn about long term digital preservation
  - Gain access to a user community and data
  - Evaluate technology for preservation
  - Apply concepts to IBM technology



http://www.casparpreserves.eu/ -- http://www.haifa.il.ibm.com/projects/storage/datastores/caspar.html

Backup

References

Publications:

- Preservation DataStores: New Storage Paradigm for Preservation Environments“
  - IBM Journal of Research and Development on storage Technologies and Systems, Volume 52, Number 4/5, 2008
- “Preservation DataStores: Architecture for Preservation Aware Storage”
  - IEEE Conference on Mass Storage Systems and Technologies (MSST), September 2007, San Diego, USA.
- “The Need for Preservation Aware Storage - A Position Paper".
  - *ACM SIGOPS Operating Systems Review*, Special Issue on File and Storage Systems, Volume 41, Issue 1 (Jan 2007), pp 19-23.
- “Towards OAIS-Based Preservation Aware Storage - A White Paper“.
  - http://www.haifa.il.ibm.com/projects/storage/datastores/public.html

Patents:

- IL8-2008-0206: A Method for Enrichment of Preservation Objects in a Preservation System – under evaluation
- IL8-2008-0205: A Method for Automatically Creating Collections of Preservation Objects in a Preservation System – under evaluation
- IL8-2008-0044: A Method for Preservation Aware Fixity Computations – rated file
- US7356480: Method of data transformation via efficient path discovery using a digraph – issued