Persistent Preservation Challenge: Experience and Recommendations

Donald Sawyer/VIE Inc.

3 April 2009

National Space Science Data Center (NSSDC)

- 45 years of operation
  - Over 2300 digital collections from 1300+ space flight instruments on 370+ spacecraft yielding 25+Terabytes
  - Over 1000 collection-supporting documents, many originally on paper or film and often addressing multiple collections
- Dramatic changes seen
- Also expect dramatic changes over next 45 years

NSSDC Storage Media Evolution

- 7-track, 200 bpi magnetic tape, documentation on paper
- 9-track, 800 bpi magnetic tape
- 12” optical disks
- 3480 cartridges
- Floppy disks, small cartridges of various types
- CD-ROMs
- Digital Linear Tape (DLT) jukebox

NSSDC's Hardware Evolution

- PDP-11, programs on paper tape
- IBM 7094 with six 7-track drives; software/control from key-punch cards
- Replaced by MODCOMP IV with 0.5MB memory and four 50MB drives
- VAX 11/780, VAX 8650, VAX 6410, VAX 9410, MicroVAXes
- MODCOMP Classic
- Britton-Lee IDM 700 data base server
- DEC, SUN, and SGI UNIX workstation

NSSDC's System Evolution

- Received media became storage media
  - Duplication for security and for distribution
  - Software written to manage digital information about data and operations
- Many file types established to hold different categories of management information
- Electronic data distribution service using relational database, staged data to disk for anonymous FTP pickup
- Established OAIS Archival Information Package (AIP) implementation; moved from VAX to UNIX for major processing and storage functions
- Currently moving all data to AIP and writing to DLTs

NSSDC's Process Evolution

- Well documented manual procedures for information entry
  - Key-punch operators digitized information from paper forms
- Dump of received data in hex characters for manual comparison with documentation
- Migration from 7-track to 9-track to 3480 using copy and stacking to higher density media
- Adopted OAIS concepts for migration, preservation, and reporting to Management

NSSDC's Evolution Impact on its Preservation Function

- Over 45 year, tremendous changes
  - Technology
  - Expectations of user communities
- Supported processes and procedures changed dramatically
- Should not be surprising some digital information lost or corrupted
  - Some additional loss probably unknown until files requested by users
- Fundamental reason?

The Persistent Preservation Challenge

- Preservation systems sometimes (and eventually will) APPEAR to function properly while ACTUALLY loosing or corrupting information
  - Human data-entry errors
  - Hardware/software errors
  - Organizational response to Workload versus Resource Mismatch
- Expect same challenge over next 45 years!

Human Data-Entry Errors

- NSSDC found human data entry errors of 1-3%
  - Key punch verifiers used to REDUCE error rate
- Recent migration-process analysis found operator errors could sometimes result in unrecognized tape misidentification
  - Established parallel operations to REDUCE error rate
- All full service archives/repositories will have human input
  - What level of risk is acceptable?

Hardware/software Errors

- Recent Migration software also performed an analysis and repair of certain past ingest artifacts
  - Very well tested, used for 2 years, millions of AIPs
  - Nevertheless, eventually found thousands of file truncations had still occurred on one particular type of data
  - Traced to low level read routine error in underlying “VAX/VMS” operating system (WHO CAN YOU TRUST?)
- Even widely used complex software likely to have ‘bugs’ that eventually leads to hard-to-recognize information loss

Workload vs. Resource Mismatch

- NSSDC experienced times with significant budget cuts and increasing workloads
  - Operations staff adopted various ‘shortcuts’
  - Reduced fidelity to existing procedures
  - Reduce reliability of supporting information
  - Not recognized until subsequent migration years later
- Reduced double-checking of human entry is easy target to address workload/resource mismatch
- Adequate attention to detail over long periods is difficult to maintain

Recommended Mitigating Approaches

- Reduce human involvement where practical
  - Perform risk analysis on impact of inevitable human-generated errors
  - Take error reduction steps as appropriate to risk tolerance
- Attempt to add independent and automated checking of processing results
  - E.g., characterize expected results and check
- Using risk analysis, highlight to Administration and Management the potential for information loss due to workload/resource mismatches