

---

# SharePoint Archival Storage Strategies & Technologies

January 2009

---

Bud Porter-Roth  
Porter-Roth Associates

415-381-6217

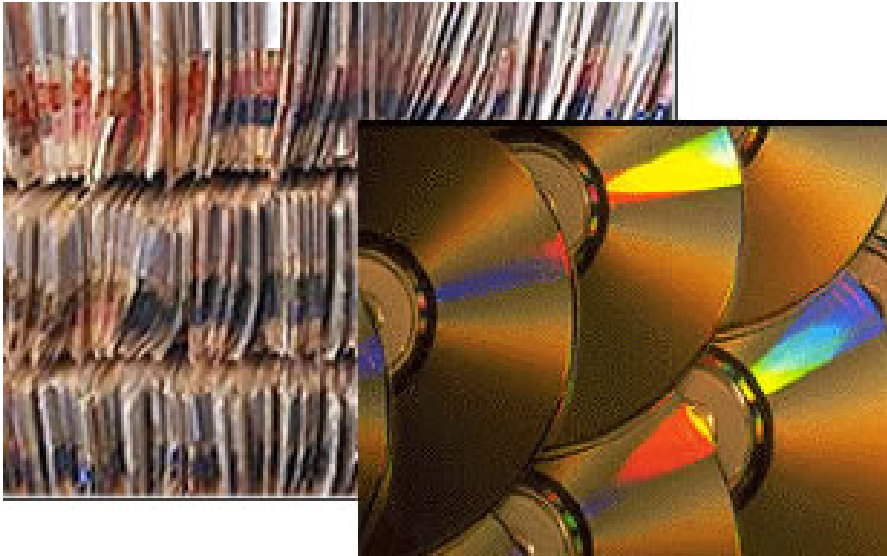
[budpr@erms.com](mailto:budpr@erms.com)

<http://www.erms.com>

# Agenda

---

- Introduction
- The Preservation Problem
- SharePoint
- Recommendations



# Introduction



Basic Need



Disaster Recovery

Compliance



# Define Archival Storage

---

- Archival storage is a requirement to store a physical or digital document longer than 15 years.....
- Archivists consider 100 years to be the baseline
- How will you store a document or library in SharePoint for 100 years? (will SharePoint still be around in 100 years? Will the version 2109 be able to read the 2007 version?)

# Drivers for Archival Storage

---

## ■ Compliance

- ✓ Length of patient's life +2 years
- ✓ SEC 17a – length of account +6 years

## ■ Legal

- ✓ Legal hold documents 1 year – end of case +??
- ✓ Some case documents stored permanently

## ■ Business Operations

- ✓ Company records – life of company + ??

# The Technical Preservation Problem

---

- The problem is actually two separate, sort of unrelated issues:
  - ✓ Hardware and software to store and read documents becomes obsolete
    - ✓ Hardware, OS, Applications, Drives – tape, disk, etc.
  - ✓ The format that the documents are in becomes obsolete or changes such that early versions are not readable
    - ✓ WordStar, VisiCalc, Word, PDF, XML, PDF-A, XPS

# The Management Preservation Problem

---

- Management may not be committed to long-term archival strategies as a necessary business procedure
- Management may be unwilling to fund long-term archival project (“I’ll be gone in 4 years...”)
- Problem is quite complex and requires a commitment to understanding the business needs over a long period of time (“I’ll be gone in 4 years...”)
- SharePoint, Lotus, eRoom, etc. encourage “unregulated” document sites and sprawl, which makes it even harder to grasp the overall problem and potential costs

# The Preservation Problem

---

## ■ The “Problem” in brief

- ✓ Software formats change and become non-supported
- ✓ Software formats fall out of favor over time and disappear
- ✓ Hardware drives change and become non-supported
- ✓ Storage media changes overtime and becomes obsolete
  - ✓ Floppy disks (5 ¼, 3 ½ )
  - ✓ Optical disks (WORM, CD, DVD)
  - ✓ Tape (many flavors of)
  - ✓ Portable storage media like the “Memory Stick” in use today

- ## ■ With all of the above issues, for digital documents, it means that there is a strong chance that you will be forced to convert/migrate something to something else over time – as a, in the foreseeable future, continuing process (\$\$\$).

# The Preservation Problem - Format

---

- TIFF (Tagged Image File Format) usually with ITG Group 4 compression (owned by Adobe, no plans for enhancement)
- JPEG (Joint Photographic Experts Group)
- GIF (Graphic Interchange Format)
- PNG (Portable Network Graphics)
- Native file formats (Word, Excel, WP etc) also known as “Born Digital” documents
- PDF, PDF/A, PDF/X, Microsoft XPS
- Many other proprietary electronic formats (especially proprietary compression algorithms)
- Paper
- Film (various types = various readers)
- Audio – (various types)

# The Preservation Problem

---

- What is the best option for preserving electronic documents over archival time spans? (Disregarding the hardware storage issues)
  - ✓ TIFF? A “digital picture” of your page
    - ✓ Widely adopted standard for document imaging
    - ✓ Not human readable without the software
    - ✓ No programmatic access to underlying text without OCR
  - ✓ XML? A format description of the page – a style sheet
    - ✓ Good for describing logical structure, but not appearance
    - ✓ Many incompatible domain-specific schemas
  - ✓ Native Format (e.g., MS Word)?
    - ✓ Several ubiquitous, but closed proprietary formats
    - ✓ Can you spell WordPerfect?
  - ✓ PDF? PDF/A?
  - ✓ Microsoft XPS?

# Desirable Properties of a Format

---

- Device independence
  - ✓ Can be reliably and consistently rendered without regard to the hardware/software platform
- Self-contained
  - ✓ Contains all resources necessary for rendering
- Self-documenting
  - ✓ Contains its own description
- Transparency
  - ✓ Amenable to direct analysis with basic tools

# Adobe PDF and PDF/A

---

- PDF is a ubiquitous open format for electronic documents
  - ✓ Proprietary, but with publicly available specification
  - ✓ Companies, other than Adobe, make PDF products
  - ✓ Has gained widespread use throughout the world
- Many statutory, regulatory, and institutional policies mandate the retention of PDF-based documents over multiple generations of technology
- The feature-rich nature of PDF can complicate preservation efforts (hence PDF/A)

# PDF/A

---

- PDF/A is intended to address three primary issues:
  - ✓ Define a file format that preserves the static visual appearance of electronic documents over time
  - ✓ Provide a framework for recording metadata about electronic documents
  - ✓ Provide a framework for defining the logical structure and semantic properties of electronic documents
  - ✓ Is an ISO Standard - ISO 19005-1:2005

# PDF/A

---

## ■ PDF/A constraints include:

- ✓ Audio and video content are forbidden
- ✓ JavaScript and executable file launches are prohibited
- ✓ All fonts must be embedded and also must be legally embeddable for unlimited, universal rendering
- ✓ Colorspaces specified in a device-independent manner
- ✓ Encryption is disallowed
- ✓ Use of standards-based metadata is mandated
- ✓ And others.....

# PDF/A ....Nevertheless

---

- PDF/A may not be the last preservation format you will use or need
- However, proper application of PDF/A should result in reliable, predictable, and unambiguous access to the full information content of electronic documents

# Microsoft XPS

---

- XPS is an abbreviation for the **XML Paper Specification**
- The XML Paper Specification describes electronic paper in a way that can be read by hardware, read by software, and read by people.
- The XML Paper Specification itself is platform independent, openly published, and available royalty-free and Microsoft has integrated XPS-based technologies into Microsoft Windows Vista operating system and the 2007 Microsoft Office system.
- <http://www.microsoft.com/whdc/xps/default.msp>

# Recap

---

- Hardware

- ✓ Disk drives

- Software

- ✓ Programs

- Format that the documents are in

- ✓ The “extension” .wp, .doc,

# But wait, we have many

---

- How many systems do you have that produce and store “potential” archival data?
  - ✓ SharePoint
  - ✓ SAP
  - ✓ Documentum, FileNet, OpenText, etc.
  - ✓ Lotus
  - ✓ eRoom

# SharePoint Archiving

---

- Documents are normally stored on magnetic disk (various configurations)
- Backups are not typically “document specific” but backup an entire Farm or site collection
- Restoration of a “library” is restoration of a complete site collection to get to the library
- Backups are typically bound to a site collection and libraries within that site collection
- Backup tapes may be recycled every xx days
- Can you depend on this combination for long-term archival storage and compliance?

# SharePoint Archiving

---

- Archival documents need to be locked such that they cannot be changed or deleted
- Archival documents may need to be separated from the original library to an archival library
- Magnetic disk/backup tape (regular cycle) may not be adequate
- May need a special backup routine for archival storage documents

# SharePoint Archiving

---

- If published and backed up to a unique archival library, ensure that document metadata remains intact
- If moved to a new storage media or location, ensure that metadata remains intact
- If restored to a SharePoint library, ensure that metadata does not change
- Ensure that documents are in a long-term format (but this is complex issue and you should work with your records management specialist)

# Recommendations

---

- This is still a wild frontier, with no certain outcome or single standard “The good thing about standards is that there are so many of them....”
- When in doubt about long-term storage of vital documents, paper or film is still a good answer (depending on volume)
- Beware of new technologies, even ones that are “standards”
- TIFF, JPEG, PDF, PDF/A are recommended
- XPS needs to be more universally adopted
- The weight of in-place document formats will mean that change will be very slow and may stop change unless a dramatic “out of the blue” technology appears (by analogy, think of PDF versus XPS)

# Recommendations

---

- Consider establishing a separate group that is specifically chartered with archival storage
- Stakeholders should include Information Management, IT, Legal, HR, and other departments with archival storage requirements
- Position should be staffed permanently (role-based) not as Jane Smith or Joe Blow

# Conclusion & Questions

---



Finally!

Questions?