PDF/A Competence Center

www.pdfa.org
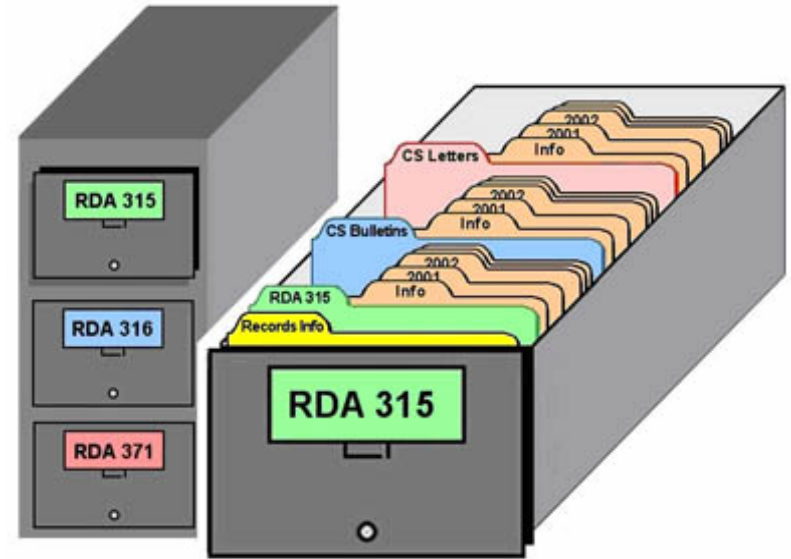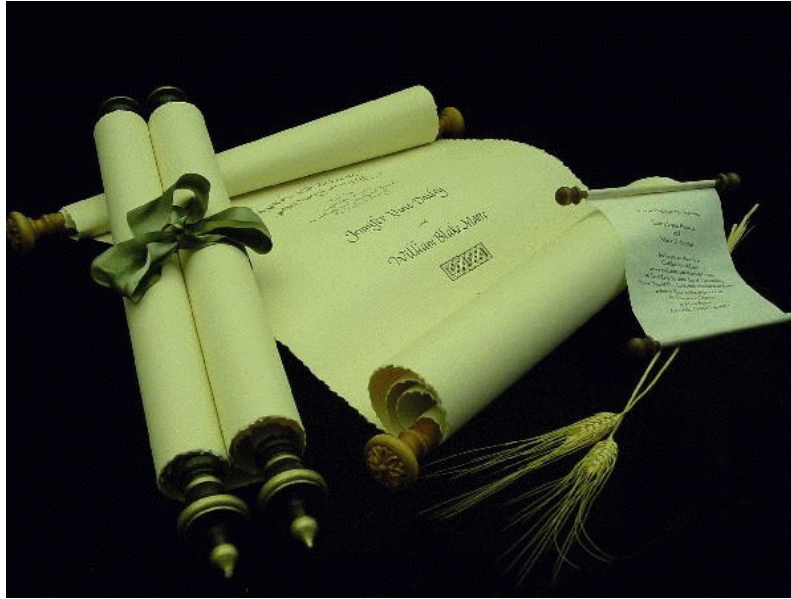
PDF/A 101:
An Introduction

10 April 2008

Stephen D. Poe, EDP
VP, Product Management
Crawford Technologies

Stephen D. Poe, EDP
Crawford Technologies

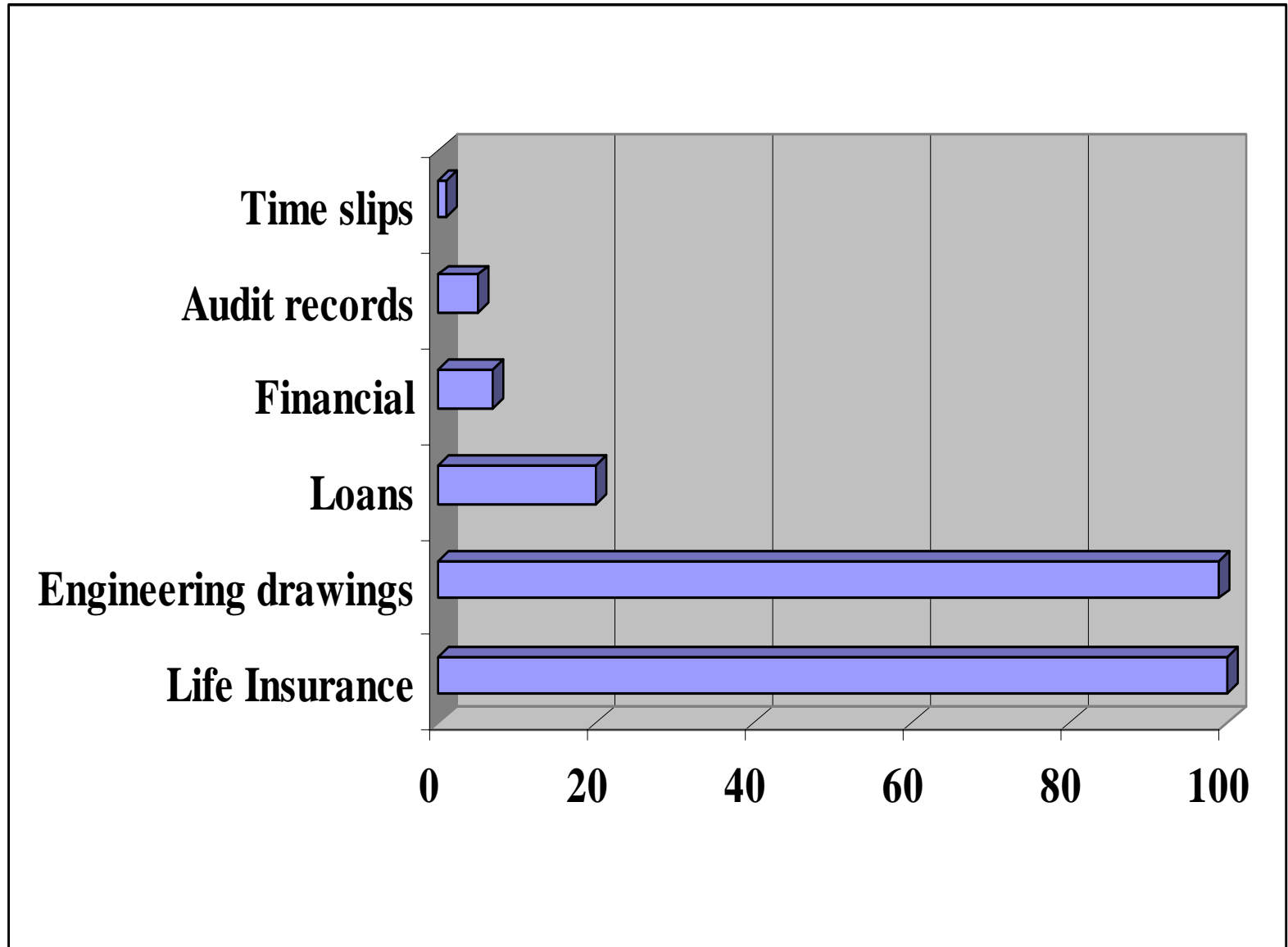# Archive Technologies

Stephen D. Poe, EDP
Crawford Technologies

Archivist Questions

- **Media Lifetime?**
  - **Paper – hundreds of years**
  - **Microfiche – dozens of years**
  - **Magnetic – perhaps a decade?**
  - **Optical – Unknown?**

- **Reader lifetime?**
  - **Paper – while language exists**
  - **Microfiche – decades**
  - **Magnetic – decades**
  - **Optical – perhaps 2-3 OS generations**

- **Key Issues in electronic archive & retrieval**
  - **Obsolete formats**
  - **Reader software is obsolete – no OS to run it on**

Legal/Regulatory Retention Periods

Stephen D. Poe, EDP
Crawford Technologies

Business Issues

- ## How do we meet:
  - ### legal and regulatory requirements
    - to hold electronic documents
    - for the mandated length of time?
      - In a cost effective manner?
      - With a defensible plan to manage them?

Aspects

- **Storage Format**

- **Retrieval/display software**

- **Archival system**
  - **HW**
  - **SW**

- **Process and procedures**

Ideal Storage Format Requirements

- **Accessible**
  - **No encryption, no proprietary formats**

- **Platform, OS, device independent**
  - **Can be read, understood and displayed on many HW/SW platforms**

- **Published specification**
  - **Open, accepted specification controlled by standards organization**

- **Self-contained**
  - **No external resources needed – including fonts**

- **Transparent**
  - **Can easily be read, parsed with non-proprietary tools**

- **Widely distributed**
  - **Accepted by both industry and governments**

Stephen D. Poe, EDP
Crawford Technologies

Candidates?

- **Raster/TIFF**
  - Broad acceptance, but obsolete
  - Loss of information – no text, structure, individual graphics
  - Creation from current systems involves throwing away information
- **Vendor Formats**
  - Proprietary formats, future unsure, not designed to be self-contained
- **XML**
  - Ability to exactly duplicate look-and-feel difficult
    - XSL/FO still not widely accepted
  - Too many DTDs, Schemas
- **PDF**
  - If tightly constrained
  - Already widely accepted

PDF/A

- **PDF Archive format specification**
  - **PDF/A**
  - **Standard, stabilized archive format**
  - **Retain exact same look-and-feel**
    - "a file format based on PDF which provides a mechanism for representing electronic documents in a manner that preserves their visual appearance over time, independent of the tools and systems used for creating, storing or rending the files."
  - **Specifies format – not archive system or process**

www.pdfa.org

PDF/A Specification

- **Effort started in 2002 by:**
  - **AIIM (Association for Information and Image Management)**
  - **NPES (National Printing Equipment Association)**
  - **Administrative Office of the U.S. Courts.**

- **ISO standard – 2005**
  - **ISO 19005-1: Document management – Electronic document file format for long-term preservation**
    - Part 1: Use of PDF 1.4 (PDF/A-1)".

- **Today**
  - **AIIM is lead on ISO Standard**
  - **PDF/A Competence Center is industry association**

PDF/A Levels (2 currently)

- ## PDF/A-1a (Level A Conformance)
  - ### Full compliance with the currently approved PDF/A Standard ISO 19005-1: Part 1
    - PDF/A-1a ensures the preservation of a document's logical structure and content text stream in natural reading order.
    - The text extraction is especially important when the document must be displayed on a mobile device (for example a PDA) or other devices in accordance with Section 508 of the US Rehabilitation Act.
    - In such cases the text must be reorganized on the limited screen size (re-flow).
  - ### This feature is also known as "Tagged PDFs".

Stephen D. Poe, EDP
Crawford Technologies

PDF/A Levels (2 currently)

- ## PDF/A-1b (Level B Conformance)
    - ### Minimal compliance to ensure that the rendered visual appearance of the file is reproducible over the long-term.
        - PDF/A-1b ensures that the text (and additional content) can be correctly displayed (e.g. on a computer monitor), but does not guarantee that extracted text will be legible or comprehensible.
        - It therefore does not guarantee compliance with US Section 508 (disability access).

PDF/A - Future

- ## PDF/A Part 1 or PDF/A-1a and PDF/A-1b
  - Based on PDF 1.4 specification

- ## PDF/A Part 2
  - In development
  - Based on selected functionality from PDF 1.5, 1.6, & 1.7
  - Backwards compatible but not forward compatible
  - Project approval January 2008
  - Current schedule is approval in 2009-2010

- ## PDF/A Part 3
  - In early design

PDF/A Internals

- ## Must be totally self-contained
  - ### No external resources, pointers, links to external content
  - ### Fonts must be included – even Acrobat Base 14

- ## Some functionality forbidden
  - ### Audio, video media inclusions
  - ### No encryption, LZW compression
  - ### Transparencies

- ## But some critical functions retained
  - ### Digital signatures
  - ### Metadata

Stephen D. Poe, EDP
Crawford Technologies

10 April 2008

Migration

- ## For existing archives
  - ### Requires conversion

- ## For new archives
  - ### Requires new processes, products
  - ### Consider parallel processes –
    - PDF for short-term archives for customer viewing
    - PDF/A for long-term archive of record

- ## Downside
  - ### PDF/A files are larger

PDF/A Products

- ## Types
  - ### PDF/A compliance verification
    - Starting with Adobe Reader 7
  - ### PDF/A creation
  - ### PDF/A conversion

- ## Target processes
  - ### Workstation (low volume)
  - ### Enterprise (high volume, production)

- ## Most only produce PDF/A -1b
  - ### To produce -1a requires a much more sophisticated production process

Archive strategy

- **PDF/A is one component**

- **Also requires:**
  - **Archive system design and implementation**
  - **Corporate processes and procedures**
  - **Detailed knowledge of what is to be archived**
    - Current production processes
    - Future production processes
    - Legacy data and documents

Summary

- **PDF/A format meets the needs for long-term archive**
  - **Functional and legal**

- **PDF/A is a format only**
  - **Also requires planning and implementation of an overall long-term archive strategy**

**Stephen D. Poe, EDP
Crawford Technologies**

For More Information

Stephen D. Poe, EDP
Vice President Product Management
Crawford Technologies
+1.214.532.0443
+1.416.410.1467
spoe@crawfordtech.com
www.crawfordtech.com