

PDF Archiving

Problem Description

Customers who intend to archive PDF documents have unique requirements to ensure that the intended document is readable for an indefinite amount of time. In order to allow PDF to meet this requirement they believe that a subset of PDF should be published and controlled by an external group who would be capable of overseeing it's evolution to ensure its longevity.

Additionally, customers need a means to defined business rules for archives and allow for simplified creation, validation and processing of the archived collection.

Requirements

1. Guidance for Archiving

- a. Provide a guidance document, which outlines the proposed strategy and its implementation.
 - i. Guidance Provides:
 1. Technical Overview
 2. Usage Overview
 3. Description of Inter-Governmental Office Adoption Path
 4. Reference FDA's guidance for Industry

2. PDF

- a. Provide a PDF version release intended for Archiving so that customer can be ensured longevity and control over PDF for this specific application.
 1. Allow for Archive independent group ownership and definition of PDF version release for Archiving
 2. Allow for Archive independent group to define acceptable fonts for archiving
 - a. Minimum Byte-Size
 - b. Technical Font Support (chemical, mathematical, genetic..)
- b. Need special fonts to represent mathematical/chemical and genetic formulae
- c. Must ensure that pagination and form are maintained and guaranteed (footnotes, references, legal legibility)
- d. Must be able to search against meta data
- e. Provide PDF Subset for archiving
 1. Include
 - a. Standards and Guidelines Specification
 - b. Open source for render and viewing (code escrow)*
 - c. Government Specific: Include process for inter-operability between Govt. agencies
 - i. Headers

* Not a hard requirement

- ii. Format
- iii. Character Sets
- iv. Electronic Sigs.
- v. GPEA, Fed. Records Act compliance
- vi. Security (DOD 5015.2 Standard)
- vii. Meta-Data (intelligent retrieval)
- viii. Audit Trail
- ix. Intelligent query from a knowledge base
- x. XML embedding and export

3. Acrobat

- a. Ability to 'Save As' PDF archive format
 - 1. Include PDF and all attachments which make up archive
 - 2. Ability to scan Archive PDF and notify user if the following unacceptable items are included:
 - a. Encryption
 - b. Proprietary compression methods (anything that isn't public domain) (although still want small files)
 - c. External references to fonts
 - d. External references to other files
 - e. Buried Executables
 - f. Proprietary fonts
- b. Ability to open an Archive PDF
 - 1. Provide notification to the user if the Archive PDF contains
 - a. Encryption
 - b. Proprietary compression methods (anything that isn't public domain) (although still want small files)
 - c. External references to fonts
 - d. External references to other files
 - e. Buried Executables
 - f. Proprietary fonts
 - 2. Provide notification to the user if the file is valid archive
 - 3. Provide notification to the user if the file is invalid archive
 - 4. Provide ability to modify file contents to meet valid archive requirements
 - a. Delete files
 - b. Apply non-proprietary compression
 - c. Remove external font references
 - d. Remove external file references
 - e. Remove proprietary fonts or fonts which are incompatible with archive format
 - f. Remove encryption (if possible)
 - g. Convert binary PDF to ascii PDF

* Not a hard requirement

- c. Ability to have batch and interactive filters for file scanning
- d. Provide a method for supporting configurable checking in the archive creation process
 - 1. Archivist Setting – Intended to check against all requirements for submission validation
 - 2. User Setting – Intended to check against all requirements for creation validation
- e. Ensure ease of use in creation process
- f. Have two modes for creation of Archive
 - 1. Enforce Valid Archive (does not allow save if file is invalid)
 - 2. Flexible (no validation)
- g. Provide templates for specific archiving file types (include fonts, layout, filters and auto-correct and check specific to the template)
 - 1. These would be created by individual agencies (not a hard requirement)
 - a. Acceptable Fonts
 - b. Acceptable Attachments
 - c. File Size

4. General Requirements

- a. Ensure that Archive Submission guidelines are International in scope
- b. Ensure that applications can read Archive document for at least fifty years
- c. Archivists must feel they have total document control upon receipt. This means that programmatic access, extraction of components and management from a user-interface.
- d. Ensure that support for Archiving is International in Scope

5. Programmatic Control

- a. Provide API for creation and validation of Archive PDF format
 - 1. Support for templates
- b. Ensure Archive documents meta data can be searched and retrieved
- c. Ensure that Archive documents can be compared for content to identify duplicates

6. Next Steps

- a. Meet at US Courts to outline PDF Archiving Proposal and determine additional requirements.