Resumen del formato PDF/A y sus variantes

Gustavo Sánchez Muñoz

(Mayo de 2013)

Hace siete años traduje y publiqué en este sitio un artículo llamado "Freguntas y Respuestas frecuentes (FAQ) sobre el estándar PDF/A", hecho por el PDF/A Competence Center.

Esta página no sustituye a dicho FAQ, que está muy bien escrito y explica las razones de la existencia de PDF/A y de sus características. Esta página sólo intenta complementar las carencias que ese FAQ tiene debido al paso del tiempo. Su objetivo es que quien la lea tenga una idea clara de qué nivel de PDF/A necesita para una tarea de archivo concreta y no se ahogue en una pequeña sopa de letras.

La finalidad de PDF/A

Archivar es guardar la información para preservarla a largo plazo. Con la digitalización de los datos, el formato PDF se ha convertido en una de las opciones favoritas de archivo digital de información. Sin embargo, como un PDF admite muchas características que dificultan la conservación a largo plazo, varios organismos y empresas se reunieron para desarrollar un estándar de PDF para la conservación y archivo de la información a largo plazo.

Decidieron qué características debía tener un PDF para archivo, cuáles no podía tener y cuales llegado el caso se permitían pero no eran obligatorias. Este conjunto de decisiones se hizo público con el nombre de PDF/A. Cualquier PDF

que reuna dichas catacterísticas y contenga unas marcas internas que lo identifique como PDF/A es un PDF/A. Ni más ni menos.

PDF/A y el mundo legal

Uno de los aspectos destacables del estándar PDF/A es que garantiza la fiabilidad e integridad de la transmisión de la información, por lo que muchos países lo han adoptado como especificación de fomato de la documentación legal digital. Ésta es la razón que ha llevado a algunas personas de ese ámbito, que no tienen mucha habilidad digital, a darse de cabezazos contra él. Tranquilos: Hacer un PDF/A desde Microsoft Word u OpenOffice es mucho más sencillo de lo que parece.

Niveles de PDF/A

Con el paso de los años, el formato PDF ha evolucionado, permitiendo cosas antes imposibles. Eso también ha ocurrido con los programas capaces de crearlos, leerlos o modificarlos. Por eso, los estándares PDF/A se han ido revisando, apareciendo nuevos niveles, cada uno con unas caracterísicas y objetivos distintos. A día de hoy —mayo de 2013— existen tres modalidades de PDF/A: PDFA-1, PDFA-2 y PDFA-3.

Un número mayor no implica un nivel *mejor* o más avanzado de PDF/A. Por eso ninguno de los estándares nuevos ha sustituido a otro anterior. La razón es que cada nivel se adapta mejor a unas necesidades de archivo para que los usuarios no se vuelvan locos reconvirtiendo sus archivos digitales de versiones viejas a nuevas.

Clases de nivel: B, A y U

Cada uno de los niveles del estándar PDF/A tiene dos o tres subniveles. El llamado subnivel "B" (de "*b*ásico") es aquel en el que se cumplen todas las especificaciones de ese estándar.

En los casos en los que existe un subnivel "U", éste es simplemente un cumplimiento básico —subnivel "B"— con el requisito añadido de que los textos del PDF tengan siempre su equivalente en codificación Unicode. Esta condición se hace para garantizar la indexación y lectura correcta de los textos.

El subnivel "A" (de "*a*ccesible") es aquel en el que además de las condiciones de los otros dos anteriores (B y U), se cumple que la estructura del contenido —especialmente de los textos— está etiquetada de forma autodescriptiva; es decir: que se describe el orden y jerarquía de la lectura.

La accesibilidad no es sólo cosa de invidentes

Si piensas que el cumplimiento de los niveles "A" es algo que sólo afecta al acceso de las personas con problemas visuales graves y que *no es para tanto* porque son pocas, estás en un error. Si aparcas el corazón por un momento —no te va a costar porque igual lo tienes pequeñito— piensa en los programas de análisis e indexación de contenido. Para que las soluciones de inteligencia artificial puedan trabajar con los textos, deben estar bien descritos y estructurados—¿cómo saber, por ejemplo cómo leer un texto multicolumna con despieces?—. Eso es lo que implica el etiquetado correcto: El uso futuro del archivo con programas de reconocimiento y análisis automatizado.

1. **PDF/A-1**

Este es el primer nivel de PDF/A. Se publicó hacia finales de 2005 —aproximádamente con la aparición de Acrobat 8—. Se describe en el estándar ISO 19005-1 *(Document management - Electronic document file format for long-term preservation - Part 1: Use of PDF 1.4 (PDF/A)).* Su finalidad principal es el archivo a largo plazo de documentos basados en páginas. Sus características principales son:

- **Prohibido**

  - **Encriptamiento:** Cualquier sistema de cifrado de datos está completamente prohibido en cualquiera de los estándares PDF/A.
  - **Compresión LZW o JPEG2000:** Ambos sistemas de compresión estan prohibidos —LZW se prohibe también en muchos otros estándares PDF—. La compresión ZIP o JPEG sí se permiten.
  - **Contenido multimedia:** No se admite la inclusión de audio, vídeo o gráficos en 3D.
  - **Código ejecutable:** Ya sea javaScript o de cualquier otro tipo como, por ejemplo, un reproductor externo.
  - **Incrustación de fuentes OpenType como tales:** las fuentes OpenType se incrustan de forma incompleta ya que el formato PDF no lo permite hasta el nivel 1.6.
  - **Transparencias o contenido alternativo:** Las transparencias son los llamados "modos de fusión" *(blending modes)* nativos del formato PDF a partir del nivel 1.4 (multiplicar, trama etc...). No se permiten. Cualquier transparencia debe ser acoplada *(flattened)*. El contenido alternativo se refiere a capas de Acrobat.

  - **Enlace a archivos externos:** No se permiten archivos enlazados necesarios para la reproducción y comprensión coherente del PDF; es decir: Debe ser un PDF autocontenido, donde todo lo necesario esté *dentro* del mismo; por ejemplo: no se permite un marco gráfico que incluya una imagen no contenida en el PDF, sino situada en un servidor externo.

    En el caso de los hipervínculos, cuando el documento se reproduce en un visor que cumple con los estándares PDF/A y lo abre en ese modo, los hipervínculos aparecen inactivos —de no hacerlo, sería contenido interactivo externo—. Si se pone el programa lector en modo de no cumplir el estándar PDF/A —lo que es posible, por ejemplo en Acrobat Reader— es muy posible que los hiperenlaces aparezcan activos.

  - **Incrustación de otros archivos:** Éste es un apartado que puede causar confusión. No quiere decir que un PDF/A no pueda llevar por ejemplo imágenes o gráficos. Quiere decir que no podemos adjuntar un archivo dentro del PDF —algo que se hace a través del

menú "`Comentarios - Herramientas comentario y marca - Adjuntar un archivo como un comentario`". En esta especificación de PDF/A eso está prohibido.

- **Obligatorio**

  - **Metadatos estandarizados:** Un PDF/A debe estar autodocumentado; es decir: Se debe identificar a si mismo usando unos metadatos estructurados con una sintaxis basada en especificaciones del formato de metadatos llamado XMP. Entre estos datos están las marcas necesarias que lo identifican como PDF/A y la variante a la que pertenecen (PDF/A-1a, PDF/A-1b, etc…).
  - **Color independiente de los dispositivos:** La reproducción del color del documento no debe depender de un aparato concreto ni de un tipo de aparatos. El color debe estar descrito mediante la inclusión de perfiles de color que permiten reproducirlo en cualquier sistema.
  - **Inclusión de las fuentes utilizadas:** Esta inclusión se puede hacer total (la fuente completa) o parcial (sólo el subjuego de caracteres utilizado).
  - **PDF nivel 1.4:** El documento debe tener ese nivel del formato PDF. No puede ser superior o inferior.

Las anotaciones de texto se permiten y los formularios también pero con muchas restricciones, especialmente en el nivel PDF/A-1.

- **PDF/A-1b**

  Este es el subnivel básico de PDF/A-1. Se cumplen todos los requisitos descritos como necesarios para un PDF/A-1.

- **PDF/A-1a**

  Este es el subnivel accesible de PDF/A-1. Además de cumplirse los criterios básicos, el texto del PDF debe estar etiquetado de forma que se desriba y conserve la estructura lógica —el orden de lectura, para entendernos— del documento original. Al hacer esto se asegura que el

contenido es accesible; es decir que se reproducirá fielmente también en el caso de usuarios con discapacidad que usan programas —como los lectores para invidentes— que deben conocer la estructura del contenido del documento para poder reproducirlo correctamente.

2. **PDF/A-2**

Esta variante del estándar PDF/A apareció en julio de 2011. Es el estándar ISO 19005-2 *(Document management – Electronic document file format for long-term preservation – Part 2: Use of ISO 32000-1 (PDF/A-2)).* Se considera una ampliación de PDF/A-1. Se basa en el nivel 1.7 del formato PDF (aparecido a partir de la publicación de Acrobat 8).

Se mejora mucho el etiquetado de la estructura del PDF para su accesibilidad e indexado, y la compresión interna de elementos.

Su finalidad principal es el archivo a largo plazo de documentos basados en páginas cuyas características originales no se pueden recoger adecuadamente con el nivel 1.4 del formato PDF y que necesitan el uso de un nivel superior, más moderno, de dicho formato. No pretende sustituir al estándar PDF/A-1, que queda para documentos que por su origen no necesitan las nuevas características. Impone compatibilidad hacia atrás —todo PDF/A-1 puede ser PDF/A-2 pero no todos los PDF/A-2 pueden convalidarse como PDF/A-1—. Sus principales características son:

- **Prohibido**
  - **Encriptamiento:** Cualquier sistema de cifrado de datos está completamente prohibido en cualquiera de los estándares PDF/A.
  - **Código ejecutable:** Ya sea javaScript o de cualquier otro tipo como, por ejemplo, un reproductor externo.

- **Contenido multimedia:** No se admite la inclusión de audio, vídeo o gráficos en 3D.
- **Enlace a archivos externos:** No se permiten archivos enlazados necesarios para la reproducción y comprensión coherente del PDF; es decir: Debe ser un PDF autocontenido, donde todo lo necesario esté *dentro* del mismo; por ejemplo: no se permite un marco gráfico que incluya una imagen no contenida en el PDF, sino situada en un servidor externo.

  En el caso de los hipervínculos, cuando el documento se reproduce en un visor que cumple con los estándares PDF/A y lo abre en ese modo, los hipervínculos aparecen inactivos —de no hacerlo, sería contenido interactivo externo—. Si se pone el programa lector en modo de no cumplir el estándar PDF/A —lo que es posible, por ejemplo en Acrobat Reader— es muy posible que los hiperenlaces aparezcan activos.

- **Obligatorio**

  - **Metadatos estandarizados:** Un PDF/A debe estar autodocumentado; es decir: Se debe identificar a si mismo usando unos metadatos estructurados con una sintaxis basada en especificaciones del formato de metadatos llamado XMP. Entre estos datos están las marcas necesarias que lo identifican como PDF/A y la variante a la que pertenecen (PDF/A-1a, PDF/A-1b, etc…).
  - **Color independiente de los dispositivos:** La reproducción del color del documento no debe depender de un aparato concreto ni de un tipo de aparatos. El color debe estar descrito mediante la inclusión de perfiles de color que permiten reproducirlo en cualquier sistema.
  - **Inclusión de las fuentes utilizadas:** Esta inclusión se puede hacer total (la fuente completa) o parcial (sólo el subjuego de caracteres utilizado).
  - **PDF nivel 1.7:** El documento debe tener ese nivel del formato PDF. No puede ser superior o inferior.

Principales diferencias de PDF/A-2 respecto al nivel PDF/A-1

La primera es que al basarse en el nivel 1.7, además de admitirse muchas más propiedades, algunas de las que ya se admitían como el etiquetado de la estructura de los textos mejoran mucho. Las principales características nuevas que se permiten son:

1. **Incrustación completa de fuentes OpenType.**
2. **Compresión JPEG2000,** aunque sólo de forma algo restringida para maximizar la compatibilidad con algunos estándares PDF/X.
3. **Transparencias o contenido alternativo**. Las transparencias son los llamados "modos de fusión" *(blending modes)* nativos del formato PDF a partir del nivel 1.4 (multiplicar, trama etc…). El contenido alternativo se refiere a capas de Acrobat.
4. **Incrustación de otros archivos.** Con la restricción de que los archivos incrustados —denominados "colecciones"— sólo pueden ser documentos PDF/A-1 o PDF/A-2.
5. **Firmas digitales** según los estándares de firma electrónica PAdES *(PDF Advanced Electronic Signatures)* definidos en el nivel 1.7 del formato PDF.

- **PDF/A-2b**

  Este es el subnivel básico de PDF/A-2. Se cumplen todos los requisitos descritos como necesarios para un PDF/A-2.

- **PDF/A-2u**

  Éste subnivel cumple las condiciones de PDF/A-2b con el añadido de que el texto debe poder extraerse o recuperarse como Unicode. No necesita incluir las etiquetas describiendo la estructura.

- **PDF/A-2a**

  Este es el subnivel más accesible de PDF/A-2. Además de cumplirse los criterios básicos, el texto del PDF debe estar etiquetado de forma que se desriba y conserve la estructura lógica —el orden de lectura, para entendernos— del documento original. Al hacer esto se asegura que el contenido es accesible; es decir que se reproducirá fielmente también en el caso de usuarios con discapacidad que usan programas —como

los lectores para invidentes— que deben conocer la estructura del contenido del documento para poder reproducirlo correctamente.

3. **PDF/A-3**

Esta variante del estándar PDF/A apareció en octubre de 2012. Es el estándar ISO 19005-3 *(Document management - Electronic document file format for long-term preservation - Part 3: Use of ISO 32000-1 with support for embedded files (PDF/A-3)).* Al igual que PDF/A-2, se basa en el nivel 1.7 del formato PDF (aparecido a partir de la publicación de Acrobat 8).

PDF/A-3 es una ampliación del estándar PDF/A-2 cuya principal diferencia es que permite la inclusión de archivos que no sean a su vez PDF/A-1 o PDF/A-2, aunque les impone ciertas restricciones (definidas a su vez en el estándar ISO 32000-1). Cualquier archivo que los cumpla puede ser lo que se llama un "archivo associado" *(associated files)* y se debe establecer una asociación explícita entre cada uno de ellos y la zona del PDF que los contiene —imagen, página o sección estructural de contenido—.

Sin entrar en las complejidades del estándar, la idea básica de los PDF/A-3 es que puedan contener colecciones de archivos que no sean a su vez estrictamente PDF/A.

- **PDF/A-3b**

  Este es el subnivel básico de PDF/A-3. Se cumplen todos los requisitos descritos como necesarios para un PDF/A-3.

- **PDF/A-3u**

Éste subnivel cumple las condiciones de PDF/A-3b con el añadido de que el texto debe poder extraerse o recuperarse como Unicode. No necesita incluir las etiquetas describiendo la estructura.

- **PDF/A-3a**

  Este es el subnivel más accesible de PDF/A-3. Además de cumplirse los criterios básicos, el texto del PDF debe estar etiquetado de forma que se desriba y conserve la estructura lógica —el orden de lectura, para entendernos— del documento original. Al hacer esto se asegura que el contenido es accesible; es decir que se reproducirá fielmente también en el caso de usuarios con discapacidad que usan programas —como los lectores para invidentes— que deben conocer la estructura del contenido del documento para poder reproducirlo correctamente.

Qué nivel de PDF/A usar

No es fácil dar una respuesta general válida para cualquier caso, pero sí se pueden hacer algunas consideraciones generales:

PDF/A-3 sólo se debe usar en casos muy específicos en los que se sepa muy bien que es necesario.

Ante la duda entre PDF/A-1 y PDF/A-2, si con el mismo esfuerzo se puede hacer el segundo estándar, es mejor hacer PDF/A-2. Si el material que se va a convertir en PDF/A no se beneficia en absoluto de las ventajas del nivel PDF/A-2 —como las transparencias nativas, por ejemplo— no tiene sentido hacer un esfuerzo extra. Si es sólo una parte del material que se va a procesar la que se beneficia de esa elección, tiene sentido elegir PDF/A-2, ya que la parte restante ni se beneficia ni se perjudica.

Por lo mismo, ante la duda entre PDF/A-2b, PDF/A-2u y PDF/A-2a, si con el mismo esfuerzo se puede hacer el tercer estándar, más accesible, es mejor hacer PDF/A-2a. Las razones ya se han explicado al hablar de los tres tipos de variantes. Es posible que el esfuerzo sea mayor, pero pudiera ser que una planificación cuidadosa del sistema de creación permitiera hacerlos sin una sobrecarga extrema. La libertad del usuario es completa en este sentido.

Si ya tenemos parte de nuestro material en una variante de PDF/A-1 y vamos a cambiar a PDF/A-2 no es en absoluto necesario reconvertir del primer estándar al

segundo. El material en PDF/A-1 puede permanecer como tal salvo que algún motivo de mucho peso nos lleve a considerarlo de otro modo —y los estándares están ideados para que eso no ocurra.

En todo caso, como el archivo a largo plazo no es una tarea de un día ni para un día, se recomienda la planificación cuidadosa del método elegido y de los estándares seleccionados, teniendo en cuenta que esas decisiones pueden evolucionar con el paso del tiempo.

Cualquier comentario es bienvenido.