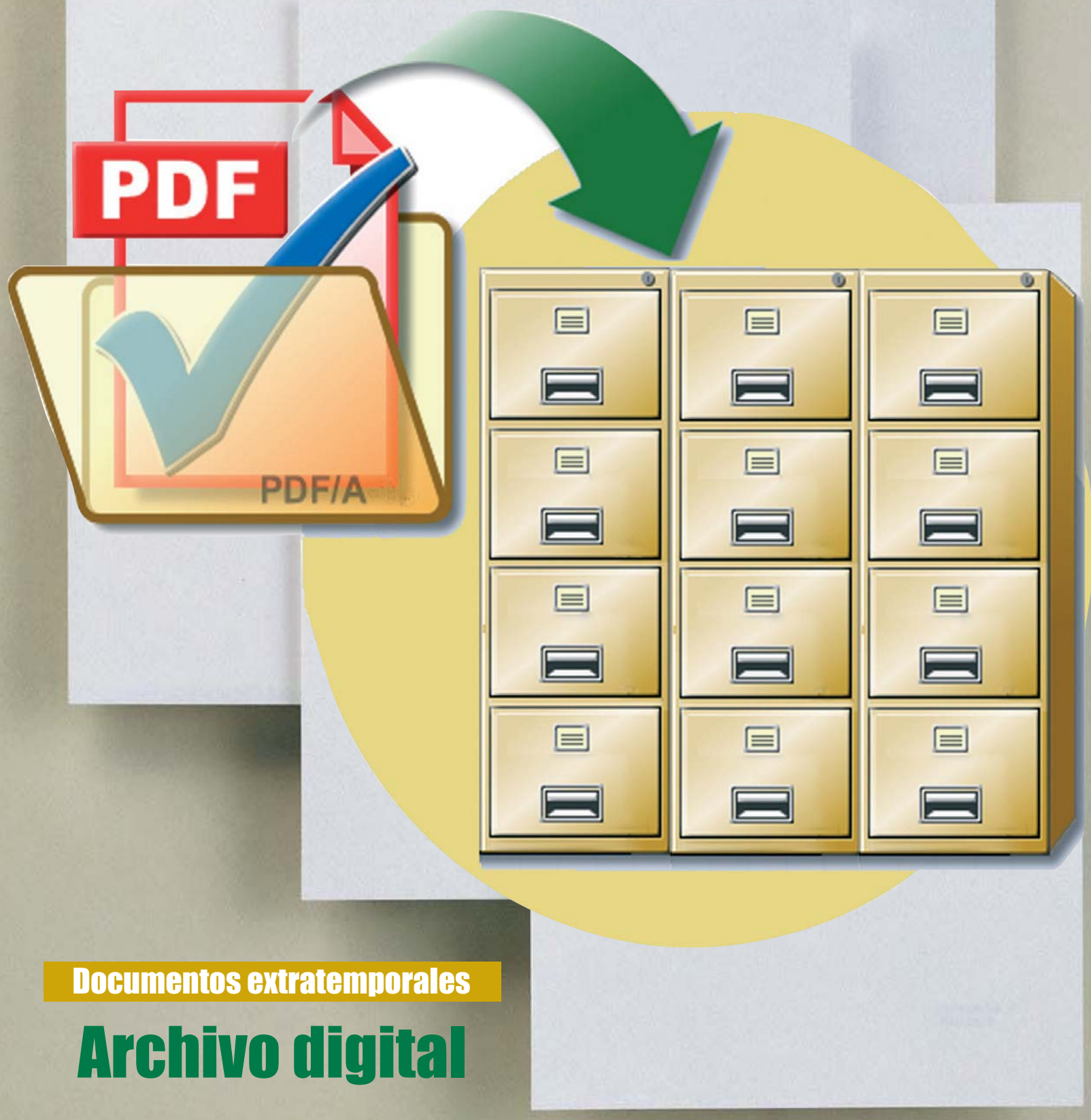
envíen

MAILING. BILLING. MANIPULADOS. GESTION DOCUMENTAL

Nº 14 ● Diciembre 2010 - Enero 2011



Documentos extratemporales

Archivo digital permanente

Introducción al PDF/A y al problema del archivado a largo plazo

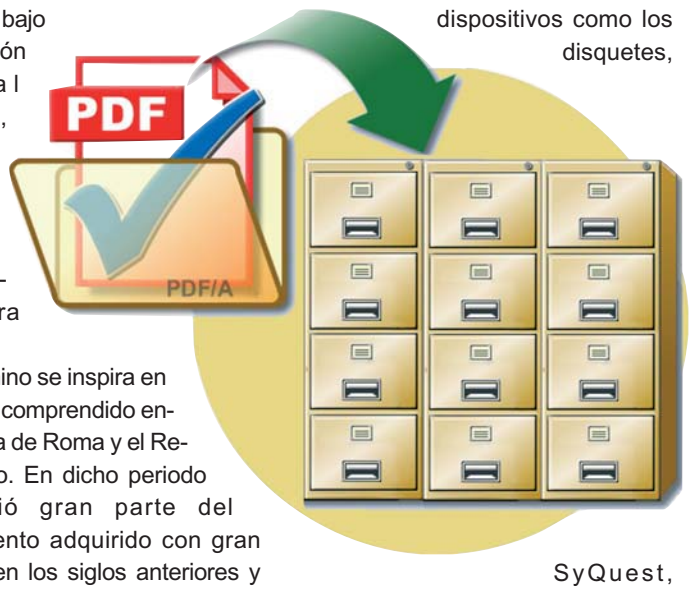
Manuel Asorey
Director de EPS y miembro del PDF/A Competence Center.

En los últimos años hemos asistido a un incremento espectacular en el volumen de archivos digitales que generamos, consumimos y almacenamos. No es solo que acostumbremos a llevar encima varios gigas de información repartidos entre lápices USB, teléfonos móviles, cámaras digitales y toda suerte de dispositivos sin los que ya no podemos vivir. Es que, además, el tamaño de los ficheros que creamos con dichos dispositivos es cada vez mayor; a modo de ejemplo, no hay más que echar un vistazo a las cámaras digitales que, incluso en sus versiones más básicas, generan ficheros de más de 10 megapíxel (alrededor de 30 megas).

Al mismo tiempo y como una de las causas de lo anterior, el precio del almacenamiento se ha reducido tanto que es muy normal encontrarnos hablando de terabytes, tamaños que antes solo las grandes empresas precisaban y/o podían pagar, incluso a nivel doméstico. Y, para hacer aún más compleja la situación, este almacenamiento está evolucionando desde el físico y local al últimamente tan en boga "almacenamiento en la nube", con múltiples servicios gratuitos y de pago compitiendo por darnos más por menos, que ha hecho que los ficheros ya no estén al alcance de nuestra mano sino en un lugar indeterminado.

Sin embargo, en paralelo a este cambio, no hemos modificado en lo sustancial nuestros hábitos. Así, incluso en estos tiempos de proliferación y sobreabundancia de medios, no hemos avanzado casi nada en las políticas de preservación. Es cierto que hay una gran diferencia entre las empresas y los particulares pero también lo es que Murphy nos visita mas a menudo de lo deseable y, en la gran parte de las ocasiones, su visita se convierte en tragedia.

Por todo lo anterior, y por las razones que expondré un poco mas adelante, desde finales de los '90 se viene hablando de la fragilidad de nuestra abundante y tecnificada información bajo la expresión "Digital Dark Age", que podríamos traducir como "Edad Media Oscura Digital".



El término se inspira en el período comprendido entre la caída de Roma y el Renacimiento. En dicho periodo se perdió gran parte del conocimiento adquirido con gran esfuerzo en los siglos anteriores y sólo gracias a los monasterios, en muchos casos gracias al trabajo de monjes que no sabían leer ni escribir pero cuyo objetivo era conservar los textos clásicos mediante la labor de copia, llegaron a nuestros días las obras griegas y latinas sobre las que se ha construido nuestra cultura occidental.

Ni que decir tiene que esta simplificada versión tiene un reflejo perfecto en nuestros días. Es cierto que los medios a nuestro alcance son mucho más amplios y que la alfabetización es general, al menos en Occidente. Pero también lo es que confiamos demasiado en la tecnología y que en muchos casos esa confianza es ciega y no basada en una reflexión profunda. A modo de ejemplo, y me pongo yo mismo en tal papel, aun conservo ficheros de mi época universitaria en dispositivos a los que es dudoso que pueda acceder, creados en aplicaciones que hace ya tiempo que han pasado a la historia. Creo que en todas nuestras mentes están dispositivos como los disquetes, SyQuest, Jazz, Zip etc. o aplicaciones como WordStar, MacWrite, WriteNow y tantas otras.

¿Soy el único con dicho problema? Evidentemente la respuesta a dicha pregunta es negativa. Incluso la todopoderosa NASA ha tenido que enfrentarse a este problema para recuperar la información del programa Viking de finales de los 70, perfectamente almacenado en cintas magnéticas pero que posiblemente solo pudo recuperarse gracias a la inversión realizada y, especialmente, gracias a que aun pudo contar con personal que había participado en dicho proyecto. Pero no hemos de confiar en que esta situación se repita en todos los casos ya que la suerte es un aliado muy poco fiable.

Una vez que el problema ha quedado claramente esbozado, y antes de entrar en el tema específico de este artículo, me gustaría hacer algunas consideraciones. No existe una receta mágica que de solución a todos los problemas planteados. Sin embargo, si podemos agrupar dichos problemas en dos categorías: problemas físicos o relacionados con el soporte de almacenamiento y problemas lógicos o relacionados con los ficheros en los que se almacena la información.

No es el objetivo de este artículo centrarse en el primer grupo ya que este es el que, quizás, tiene una solución más sencilla. Al final, y aunque asumamos que los nuevos soportes digitales son mas frágiles y tienen una vida media mucho mas corta de lo deseable, también es cierto que las bajadas de precios del almacenamiento a los que anteriormente hemos aludido hacen posible, con unos costes mas o menos razonables, tanto la rotación de soportes como la replicación de los mismos.

En cuanto al segundo grupo, es importante hacer ciertas matizaciones preliminares: el tipo de archivos digitales a conservar es demasiado amplio como para que podamos vislumbrar una única solución. A día de hoy, y al menos para quien escribe estas líneas que se declara lego en las problemáticas relativas a estos tipos de fichero, los archivos de video, audio y, en general, todos aquellos que

podemos categorizar como multimedia, presentan unas características específicas que impiden una estandarización con vistas a la preservación a largo plazo.

Documentación

Nos centraremos por ello en la documentación, un capítulo en el que se incluye la gran mayoría de los ficheros susceptibles de ser conservados tanto a nivel particular como empresarial. Y, dentro de este grupo de ficheros, distinguiremos, a su vez, entre dos tipos: los que tienen un origen completamente digital (es decir, han sido creados en el ordenador) y aquellos que han sido convertidos a digital pero cuyo origen es físico: papel. Para todos ellos existe un estándar ISO, el 19005-1:2005, también conocido como PDF/A (la A se corresponde con "Archivo"), cuya finalidad es la preservación a largo plazo manteniendo la integridad visual.

Los orígenes del PDF/A se remontan a la primavera de 2002 cuando diversas organizaciones de EE.UU., conscientes de que el formato PDF se estaba convirtiendo en el más utilizado para el almacenamiento e intercambio de ficheros electrónicos, crearon un grupo de trabajo cuyo fin era buscar formulas para garantizar que esos ficheros PDF pudieran ser visualizados de forma fiel en un futuro. Los trabajos de este grupo pronto se encuadraron dentro de ISO y, como consecuencia, terminaron convirtiéndose en el estándar para archivo.

El PDF/A es básicamente un subconjunto de la especificación PDF (en la actualidad también un estándar ISO, publicado con la identificación 32.000-1:2008) que limita a esta a fin de garantizar el almacenamiento y la visualización fiel en el tiempo.

La elección del formato PDF era, en cierta medida, lógica puesto que aportaba, al margen de su amplia aceptación y difusión, algunas características básicas para el fin propuesto: independencia de la aplicación creadora y de la plataforma (Sistema Operativo) e inclusión de todos los elementos necesarios para su representación visual.

Tenemos, entonces, que el PDF/A solventaba de un plumazo algunos de los problemas planteados anteriormente: permitía que un fichero hipotético, creado en una aplicación de procesador de texto en los años 80 y convertido a PDF/A, pudiera ser abierto en, pongamos por caso, un ordenador actual con SO Linux y, lo que es aún mejor, visualizado de la misma forma en la que lo hizo el creador del mismo. Así pues, parece que el PDF/A constituye un importante paso adelante en la búsqueda de un remedio para esa amedrantadora "edad oscura digital".

Los orígenes del PDF/A se remontan a la primavera de 2002 cuando diversas organizaciones de EE.UU., conscientes de que el formato PDF se estaba convirtiendo en el más utilizado para el almacenamiento e intercambio de ficheros electrónicos, crearon un grupo de trabajo cuyo fin era buscar formulas para garantizar que esos ficheros PDF pudieran ser visualizados de forma fiel en un futuro.

Pero, llegados a este punto es lícito preguntarse las razones por las que este formato es una mejor solución que otros en uso desde hace tiempo o, incluso, que su "progenitor", el PDF.

Si lo comparamos con el PDF, que como mencionaba anteriormente, es, desde el 2008, un estándar ISO al haber sido cedido por su creador Adobe Systems, el PDF/A no es a primera vista muy diferente. De hecho, un PDF/A es un PDF completamente válido y conforme a norma. Pero, sin embargo, un PDF ISO 32.000 no es necesariamente un PDF/A válido.

La razón para dicha circunstancia radica no tanto en lo que tienen en común ambas especificaciones como en las restricciones que el último impone al primero. Esas restricciones, a fin de no extenderme o hacer compleja la explicación, se puede clasificar en dos tipos: requerimientos y prohibiciones. A fin de clarificar estos conceptos, pongamos un ejemplo de cada uno.

Uno de los requerimientos mínimos de un PDF/A es que las tipografías utilizadas en el estén incrustadas. Dicho requerimiento no existe en un PDF por lo que es perfectamente lícito el crear ficheros en los que las tipografías no estén presentes. Pero dado que ello conllevaría una posible alteración de la representación de dicho documento, el que en principio es un fichero PDF válido no podrá ser un PDF/A conforme a norma.

En lo que se refiere a prohibiciones, es perfectamente posible en un PDF exigir una contraseña para su apertura y visualización. Pero, por razones obvias, esta característica esta terminantemente prohibida en un PDF/A ya que no podemos garantizar que dicha contraseña sea conocida en el futuro. Imaginemos por un instante que ese fichero fuese el plano de un puente y que necesitásemos acceder a el dentro de, pongamos por caso, 50 años. ¿Sería asumible el poder acceder al fichero solo para darse cuenta de que no es posible visualizar su información debido a que la contraseña hace ya tiempo que se olvidó?

El formato TIFF-G4

Pero el PDF no es el único formato de archivo posible. Dado que no quiero extenderme en demasía sobre este particular, me centraré en un formato comúnmente utilizado para archivo, el formato TIFF-G4.

Desde mi punto de vista, el TIFF-G4 no es una solución correcta para el almacenamiento y preservación de nuestra documentación. Lo primero que podemos decir de él es que, pese a lo que algunos piensen, no se trata de un formato estándar si bien hay que admitir que su uso está muy extendido, especialmente dentro de las empresas. Pero el hecho de no ser un estándar no es la peor de sus facetas. Al igual que otros formatos raster (compuestos por píxeles al igual que las fotografías) más modernos como el JPEG, su primera limitación es su propio carácter de "fotografía" de un original. Esto no parece inicialmente negativo puesto que si es una fotografía tenemos cubierto el aspecto de la representación fidedigna. Pero, como veremos, esta ventaja es extremadamente pequeña si la comparamos con las desventajas. Entre ellas debemos mencionar un mayor tamaño, la inexistencia del concepto documento (a diferencia del PDF, un fichero de 20 páginas serían, en TIFF-G4, 20 ficheros correspondientes con cada una de las páginas del mismo), la imposibilidad de incrustar los resultados de un proceso de OCR (Optical Character Recognition o Reconocimiento Óptico de Caracteres) y, por tanto, la dificultad para hacer una búsqueda por contenidos, la limitación al Blanco y Negro, etc.

Dejo en el tintero otros formatos de archivo habituales como los de Microsoft Word que, pese a los pasos en al dirección correcta dados por el gigante de Redmond, no cumplen con los requisitos mínimos esbozados anteriormente: no son multiplataforma, no garantizan la integridad visual, etc.

Ficheros PDF/A correctos

Llegados a este punto y si, al igual que yo, el lector esta convencido de las bondades del PDF/A, la siguiente pregunta a formularse es la correspondiente a cómo crear ficheros PDF/A correctos. Una pregunta tan simple no tiene, sin embargo, una respuesta igual de sencilla. De una forma rápida, y tal y

como avanzábamos anteriormente, podríamos responder que la forma de crear PDF/A correctos depende del origen de nuestros archivos: digital o analógico (papel).

En el caso del origen analógico, las ventajas de estandarizar a PDF/A son múltiples. Por un lado, nos liberamos de gran cantidad de espacio y consecuentemente nos ahorramos dinero puesto que dicho espacio puede ser destinado a actividades generadoras de negocio; por otro, obtenemos un fichero que permite, si se ha generado un PDF/A con OCR, la búsqueda por contenidos, algo que estoy seguro convendremos que incrementa enormemente la productividad al reducir los tiempos necesarios para localizar un documento dado.

Al margen de estas ventajas, cuando tenemos un origen analógico el primer paso es siempre la labor de digitalización, una labor que nos facilitará bien un formato raster como un TIFF o JPEG por cada página bien, en los modernos escáneres, un PDF del documento. A partir de este momento, el único proceso necesario sería la conversión a PDF/A y, eventualmente, el OCR del documento mediante alguna de las múltiples aplicaciones disponibles en el mercado. El resultado final sería un fichero PDF que contendría una capa de OCR y que podría ser catalogado en cualquiera de los múltiples gestores de contenidos al uso. Es importante señalar que el fichero PDF/A resultante de este proceso es mucho más reducido en cuanto a tamaño que los ficheros raster originales e incluso, en el caso mencionado de los escáneres mas modernos, que el PDF generado de la digitalización. Esto es debido a la aplicación de técnicas de compresión optimizadas para el archivado. Y, con ello, obtenemos otros ahorros significativos al reducirse el espacio necesario para almacenar estos ficheros, la necesidad de ancho de banda y tiempo de transmisión en el caso de precisar compartir los ficheros resultantes, y, finalmente, debido a su tamaño reducido, reducirse los tiempos y requerimientos de ordenador a la hora de necesitar realizar consultas. Sirva como ejemplo un caso al que nos hemos tenido que enfrentar recientemente: un documento cuyo tamaño una vez escaneado era de 102,4 MB en TIFF o de 101,04 en PDF, pasó a tener un tamaño de 3,2 MB en PDF/A. Y dicho PDF/A, al que se había añadido un proceso de OCR, permitía la búsqueda por contenidos, algo que era imposible en los originales.

Por su lado, cuando el origen de nuestros ficheros es completamente digital hemos de realizar una diferencia entre aquellos ficheros en formato nativo (es decir, en el formato de una aplicación concreta como pueda ser el caso de Microsoft Word), o en el caso de aquellos que ya son PDF.

En el caso de los ficheros nativos, el primer proceso necesario es la conversión a PDF/A, conversión que se puede hacer en uno o dos pasos dependiendo de las herramientas con las que contemos para dicho proceso.

En un paso nos referimos a la creación de PDF/A directa desde la aplicación. El caso de los ficheros de Microsoft Office, netamente de Microsoft Word puede ser un buen ejemplo ya que la propia Microsoft ha puesto a disposición de los usuarios de Office 2007 un módulo gratuito que permite guardar directamente en PDF/A.

Inicialmente parece, pues, que esta es la mejor solución. Ahora bien, hemos de ser cautos con esta aproximación puesto que el que una aplicación permita guardar sus archivos como PDF/A no quiere decir que el resultado sea conforme a norma.

En el ejemplo utilizado anteriormente, los ficheros así creados como PDF/A no son en realidad conformes a norma puesto que hay un error en el módulo de conversión que impide que sean correctos. Justo es decir que ese error ha sido ya subsanado en la versión 2010.

Validación o certificación

Independientemente del ejemplo, la creación de PDF/A desde ficheros nativos nos lleva a la necesidad de un proceso adicional: la validación o certificación.

PDF/A Competence Center, entre otras labores, tiene como finalidad última la difusión del formato PDF/A. Para ello, además de diferentes actividades como libros, seminarios, documentación, etc., ha creado una batería de comprobaciones denominada Isartor cuyo objetivo es permitir a los usuarios finales de PDF comprobar que las aplicaciones que utilizan para verificar como PDF/A hacen realmente su trabajo.

En realidad, la necesidad de este proceso es algo lógico y a lo que estamos acostumbrados en otros ámbitos. Así, por ejemplo, cuando presentamos un escrito a una instancia oficial y adjuntamos una copia de, pongamos por caso, nuestro documento de identidad, lo normal es que se realice una labor de compulsa del mismo. Dicha labor pretende garantizar que la copia y el original sean un reflejo fiel el uno del otro. Nada diferente de lo que es necesario en el caso de un archivo construido alrededor del PDF/A. El origen de los documentos puede ser diverso, como hemos visto, y la conversión a PDF/A puede realizarse mediante variadas herramientas. Así pues, es más que necesaria la labor de un "notario" virtual que garantice que los ficheros a conservar son conformes a norma y, por tanto, no contienen problemas que en el futuro impidan su utilización y/o visualización.

En este punto, al igual que en el caso de la conversión de ficheros nativos a PDF/A directamente, existen diversas herramientas en el mercado. Cada una de ellas presenta puntos fuertes y débiles, por lo que una labor importante antes de elegir al "notario" es comprobar su fiabilidad.

PDF/A Competence Center

Y en este punto entra en el juego el PDF/A Competence Center. Entre otras labores, esta organización tiene como finalidad última la difusión del formato PDF/A. Para ello, además de diferentes actividades como libros (su libro sobre PDF/A titulado "Introducción al PDF/A" es uno, si no el único, sobre el particular disponible en castellano), seminarios, documentación, etc., ha creado una batería de comprobaciones denominada Isartor cuyo objetivo es permitir a los usuarios finales de PDF comprobar que las aplicaciones que utilizan para verificar como PDF/A hacen realmente su trabajo.

Conclusión

Concluimos así este artículo en el que creemos queda esbozada una rápida introducción al PDF/A, a su fin último y a las ventajas del mismo. Es seguro que se quedan múltiples aspectos en el tintero tanto aspectos por mencionar como aspectos por desarrollar en más detalle. Pero esto no es necesariamente malo ya que en ningún momento se pretendió el realizar algo diferente de una introducción al PDF/A y al problema del archivado a largo plazo. La idea del articulo ha sido siempre no caer en contenidos técnicos sino plantear interrogantes al lector, hacerle consciente de un problema, del que muchas veces no somos conscientes, y de la solución que aporta al mismo el estándar PDF/A. Tiempo habrá para otros artículos en detalle. Entretanto, y si hay algún concepto que precise aclaración, estaré encantado de responder a las preguntas que me envíen a manuel.asorey@pdfa.org