

Formatos XML Abiertos de Microsoft® Office

Nuevos formatos de archivo para Office 2007

Publicado: Junio 2005

Actualizado: Mayo 2006

La información más actualizada en

<http://www.microsoft.com/spain/Office/preview/default.aspx>

y <http://www.microsoft.es/servidores/interop>



CONTENIDOS

Presentación de los Formatos XML Abiertos de Microsoft Office	1
Estructura de archivo	3
Paquete ZIP	3
Partes	5
Relaciones	6
Archivos con Macros y sin Macros	8
Extensiones de archivo	9
.....	9
Desarrollo de soluciones	11
Interoperabilidad de los datos	12
Manipulación de contenidos	13
Compartir y reutilizar contenidos	15
Ensamblado de documentos	16
Seguridad de los documentos	18
Gestión de información sensible	18
Formato de documentos	19
Catalogación de documentos (“profiling”)	20
Conclusión	21
Referencias.....	¡Error! Marcador no definido.

Presentación de los Formatos XML Abiertos de Microsoft Office

Por defecto, los documentos creados con Microsoft Office 2007 se basarán en una definición de formato de archivo XML. Este nuevo formato es diferente del formato de archivo binario que se ha ido manteniendo a lo largo de las versiones precedentes. El nuevo Formato XML Abierto de Microsoft Office introduce una serie de ventajas que beneficiarán no solo a los desarrolladores y las soluciones que podrán crear, sino a los usuarios individuales y organizaciones de todo tipo. Estas son algunas de las principales ventajas que reporta el Formato XML Abierto:

- **Abierto y exento de royalties** – El Formato XML Abierto se basa en las tecnologías XML y ZIP, y por tanto son accesibles de forma general. La especificación del formato y esquemas se publicará y se hará disponible bajo licencia exenta de royalties y con una cláusula de “acuerdo de renuncia a reclamación” que asegura que los desarrolladores podrán utilizar los formatos bajo la modalidad de licencia que prefieran.
- **Estándar de interoperabilidad** – Con el estándar XML como base del Formato XML Abierto, todo el proceso de intercambio de datos entre aplicaciones de Microsoft Office y los sistemas corporativos empresariales se simplifica enormemente. Al no necesitarse acceso a aplicaciones Office, las soluciones pueden alterar la información contenida dentro de un documento de office, o incluso crearlo por completo desde cero utilizando las tecnologías y herramientas estándar para manipular XML.
- **Compatible y estable** – El Formato XML Abierto está diseñado para ofrecer más robustez que los formatos binarios y con ello reducir el riesgo de pérdida de información debido a la corrupción o daño en los archivos. Incluso los documentos creados o modificados fuera de Office son menos sensibles a la corrupción, ya que los programas Office están preparados para recuperar los documentos, mejorando su fiabilidad gracias al uso de este nuevo formato.
- **Eficiente** – El Formato XML Abierto utiliza la tecnología ZIP de compresión para guardar los documentos. Este tipo de compresión de archivos debe redundar en un ahorro de costes al reducir el espacio de disco necesario para almacenar los archivos, y el ancho de banda necesario para transportarlos mediante correo, sobre la red o en Internet.

- **Seguro** – En su condición de formato de libre acceso, el Formato XML Abierto se traduce en archivos más seguros y transparentes. Los documentos se pueden compartir garantizando su confidencialidad, ya que la información identificable como de tipo personal o sensible (p.ej. nombres de usuarios, comentarios, rutas de archivos, etc.) se puede identificar con facilidad y eliminarse. De la misma forma, ciertos archivos con contenidos como objetos OLE o código VBA (Visual Basic® for Applications) se pueden identificar rápidamente para darles el tratamiento adecuado al tipo de información que contienen (código, gráficos, macros, etc.).

Puede ampliar información sobre el Formato XML Abierto de Microsoft Office y las ventajas que reporta en el sitio Web de presentación preliminar de Microsoft Office 2007, indicado en la sección de referencias al final de este documento. El resto del contenido en este Whitepaper se centrará en un análisis técnico del Formato XML Abierto y las oportunidades que ofrece a los desarrolladores.

Estructura de archivo

- *Contenedor ZIP con compresión*
- *Múltiples partes XML describiendo los datos, metadatos y datos del usuario dentro del archivo*
- *Partes no-XML soportadas como archivos nativos (imágenes objetos OLE)*
- *La estructura del archivo viene definida por las relaciones*

En el centro mismo del Formato XML Abierto de Microsoft Office está el uso de esquemas de referencia XML y el contenedor ZIP. La combinación de XML y ZIP hace posible un formato muy robusto y modular que da pie a numerosos escenarios nuevos.

Cada archivo se compone de una colección de un número indeterminado de partes. Esta colección define el documento. Las partes del documento se mantienen unidas dentro del archivo contenedor o paquete, usando el formato ZIP estándar. La mayoría de las partes son archivos sencillos XML que describen los datos de aplicación, metadatos e incluso los datos que los usuarios guardan dentro del archivo contenedor. Otras partes no-XML pueden aparecer también dentro del contenedor, por ejemplo archivos binarios (imágenes, objetos OLE) embebidos dentro del documento. Las partes pueden especificar una relación con otras partes. Este diseño es el que define la estructura de un archivo de Office. Aunque las partes son las que albergan los verdaderos contenidos del archivo, las relaciones describen cómo operan estas partes en un conjunto organizado.

El resultado es un formato de archivo XML para los documentos de Office, estrechamente integrado, pero modular y muy flexible. En las siguientes secciones vamos a explorar cada componente del Formato XML Abierto con más detalle y también se mostrará cómo se utiliza en cada uno de los tres programas de Office que disponen de estos formatos.

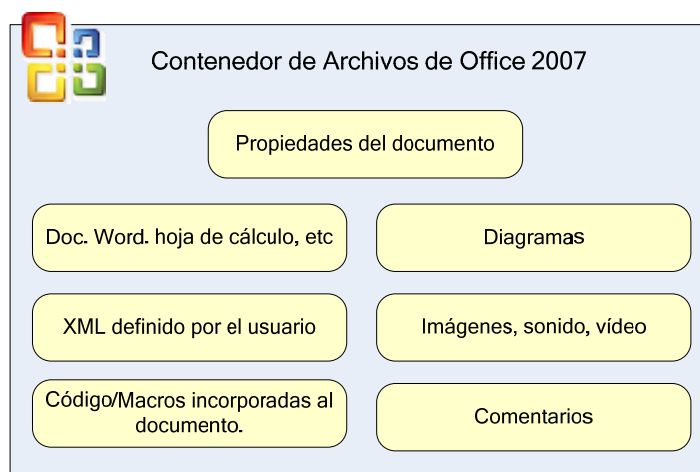
Paquete ZIP

- *Es un estándar de la industria*
- *La experiencia del usuario es la de un archivo único*
- *La compresión reduce las necesidades de almacenamiento*
- *Los desarrolladores pueden procesar el archivo con herramientas estándar*

Son muchos los elementos que intervienen a la hora de crear un documento Office. Algunos de ellos están compartidos por todas las aplicaciones de Office, como las propiedades del documento, estilos, diagramas, hipervínculos, comentarios y anotaciones. Otros elementos

son específicos de cada aplicación, como las hojas de cálculo en Excel, las diapositivas en PowerPoint o las cabeceras y pies de página en Word.

Cuando los usuarios guardan un documento con la versión más reciente o las anteriores versiones de Microsoft Office, se escribe un único archivo a disco, que después puede abrirse con total facilidad. Esta metáfora es importante a la hora de entender cómo se guarda, almacenan y comparten los documentos en situaciones reales. Al empaquetar todas las partes componentes de un archivo Office 2007 dentro de un contenedor ZIP, los documentos siguen formando parte de una única instancia de archivo. El uso de un paquete único para representar una única entidad de documento supone que los usuarios tendrán la misma experiencia que hoy cuando guarden y abran documentos con Office 2007.



Con las anteriores versiones de Office, los desarrolladores que querían manipular el contenido de un documento Office tenían que saber leerlo y escribirlo de acuerdo con el almacenamiento estructurado definido dentro del archivo binario. EL proceso sabemos que es complejo y laborioso, en gran parte debido a que los formatos de archivo binario de Office estaban pensados para accederse básicamente desde programas de Office. Los formatos se habían estructurado para reflejar al máximo posible las estructuras de memoria de las aplicaciones y poder correr en máquinas con poca cantidad de memoria, y con discos duros lentos. La alteración de los archivos binarios de Office mediante programación sin intervención de las aplicaciones Office solía ser una de las causas más frecuentes de corrupción de los archivos, y ha disuadido a muchos programadores de intentar siquiera realizar cambios sobre los contenidos de los archivos.

Se eligió el formato ZIP debido a que es un estándar muy difundido y conocido. Hay muchas herramientas disponibles para poder trabajar con el formato ZIP, que además ofrece una estructura flexible y modular que permite ampliar sus funcionalidades. Por ello los desarrolladores tendrán acceso pleno a todo el contenido de los documentos de Office 2007 utilizando algunas de las numerosas herramientas y tecnologías que funcionan con los archivos ZIP estándar. Una vez abierto un contenedor, los programadores pueden manipular cualquiera de las partes del documento que se encuentran dentro del paquete que define el documento. Por ejemplo, un desarrollador puede abrir un documento Word que utiliza el Formato XML Abierto, localizar la parte XML que representa el cuerpo del documento Word, modificar esa parte utilizando cualquier tecnología que pueda editar XML y devolver esa parte al paquete contenedor para originar un documento Office actualizado. Este escenario es solo uno de los innumerables ejemplos que serán posibles gracias al nuevo formato.

Partes

- *Son piezas modulares que componen un archivo de Office*
- *Cada parte es, básicamente, un “archivo”*
- *En formato XML en principio*
- *Se pueden utilizar otros formatos nativos para diferentes fines (imágenes, objetos OLE)*

Dentro de un paquete en Formato XML Abierto podemos encontrar muchos elementos almacenados como partes individuales. Esta modularidad es una de las características más notables del nuevo formato de archivo. La modularidad permite que los desarrolladores localicen rápidamente una parte concreta y operen directamente sobre ella. Las partes se pueden editar, intercambiar o borrar incluso, dependiendo de lo que se quiera conseguir para cubrir alguna necesidad de negocio

Ciertas partes están compartidas por todos los programas de Office, como las imágenes en miniatura, metadatos y partes de relaciones. Otras aparecen continuamente en todos los archivos como partes concretas, como las propiedades del documento. Otras muchas partes, no obstante, son exclusivas de cada tipo de documento y de la aplicación que representan. Por ejemplo, una parte “hoja de cálculo” solo la podremos encontrar en un documento Excel, y la parte “documento maestro” de una presentación solo aparece en un documento PowerPoint.

Las partes pueden estar compuestas por distintos contenidos físicos. Las utilizadas para describir los datos de un programa Office se guardan como XML. Estas partes asumen el

esquema de referencia XML, que define la característica u objeto de Office asociado. Por ejemplo, dentro de un archivo Excel, los datos que representan a una hoja de cálculo se encuentran en una parte XML compatible con el esquema de Office para una hoja de cálculo Excel. Además, si hubiesen múltiples hojas de cálculo en un libro Excel, habría una parte XML correspondiente, guardada dentro del archivo de paquete para cada hoja de cálculo. Todos los esquemas que representan parte de documentos Office estarán plenamente documentados y a disposición del público en licencia de uso exenta de royalties. Así, mediante el uso de cualquier tecnología estándar basada en XML, los desarrolladores pueden aplicar su conocimiento de los esquemas Office para interpretar y crear fácilmente documentos Office 2007.

En muchas ocasiones es preferible guardar las partes en sus tipos de contenido nativos. Estas partes no se guardan como XML. Las imágenes dentro de un documento Office, por ejemplo se guardan como archivos binarios (.png, .jpg, etc..) dentro del paquete del documento. Por tanto, se puede abrir el paquete contenedor utilizando alguna herramienta ZIP e inmediatamente podremos ver, editar o sustituir la imagen en su formato nativo. Esta técnica de almacenamiento no solamente es más accesible, sino que exige menos procesamiento interno y espacio en disco que el guardar la imagen codificada como XML. Otras partes interesantes guardadas como archivos binarios son los proyectos VBA y los objetos OLE embebidos. Para los desarrolladores, la accesibilidad hace más atractivos muchos escenarios de uso. Por ejemplo, podemos crear una solución que acceda a una colección de documentos Office 2007 y actualice un objeto OLE incrustado con una nueva versión. Esta idea y muchas otras pueden llevarse a la práctica sin tener que utilizar el programa Office o alterar el XML específico del documento.

Relaciones

- *Son partes que describen la conexión entre otras dos partes.*
- *Las conexiones se describen mediante XML*
- *Definen la estructura del formato de archivo con una fácil navegación*
- *Pueden referenciar recursos externos, vinculados.*

Aunque las partes son elementos individuales que componen un documento Office, las relaciones son el procedimiento utilizado para definir cómo se interrelacionan las partes para crear un documento. Las relaciones se definen mediante XML, que especifica la conexión entre una parte origen y un recurso de destino. Por ejemplo, la conexión entre una diapositiva y una imagen que aparece dentro de ella se identifica mediante una relación. Las relaciones

se guardan ellas mismas dentro de partes XML o “partes de relaciones” en el contenedor del documento. Si una parte de origen mantiene múltiples relaciones, todas las demás se enumeran en la misma parte de relaciones XML.

Las relaciones juegan un papel esencial en el Formato XML Abierto y cada parte es referenciada por, como mínimo, una relación. La implementación de relaciones hace posible que en el interior de las partes no haya referencias directas a otras partes, y las conexiones entre ellas se descubran directamente sin tener que analizar su contenido interno. Ya dentro de las partes, todas las referencias a relaciones se representan con un ID específico que permite mantener su independencia del esquema propio del contenido.

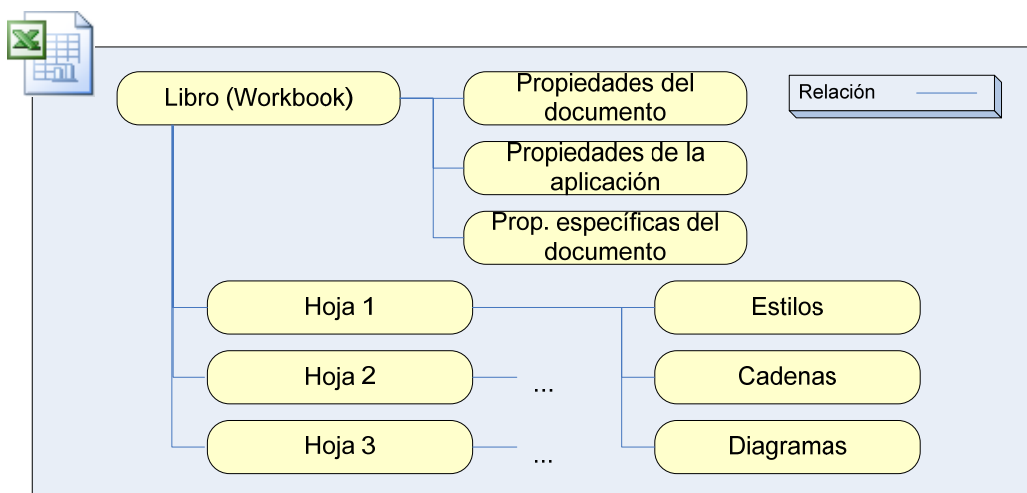


Diagrama de alto nivel de relaciones en un libro de Excel 2007

A continuación mostramos un ejemplo de una parte de relaciones en un libro Excel 2007 que contiene dos hojas de cálculo:

```
<Relationships
xmlns="http://schemas.microsoft.com/package/2005/06/relationships">
<Relationship ID="rId3"
  Type="http://schemas.microsoft.com/office/2005/8/relationships/xlStyles"
  Target="styles.xml"/>
<Relationship ID="rId2"
  Type="http://schemas.microsoft.com/office/2005/8/relationships/xlWorksheet"
  Target="worksheets/Sheet2.xml"/>
<Relationship ID="rId1"
  Type="http://schemas.microsoft.com/office/2005/8/relationships/xlWorksheet"
  Target="worksheets/Sheet1.xml"/>
<Relationship ID="rId5"
  Type="http://schemas.microsoft.com/office/2005/8/relationships/xlMetadata"
  Target="metadata.xml"/>
<Relationship ID="rId4"
```

```
Type="http://schemas.microsoft.com/office/2005/8/relationships/xlSharedStrings"  
Target="strings.xml"/>  
</Relationships>
```

Hay que destacar que las relaciones representan no solo las referencias internas del documento, sino también los recursos externos. Por ejemplo, si un documento contiene enlaces a imágenes u objetos, quedan representados mediante relaciones. Con ello los enlaces a recursos externos dentro de un documento son fáciles de localizar, analizar y modificar. A los desarrolladores les permite reparar enlaces externos rotos, validar orígenes de datos desconocidos o eliminar enlaces potencialmente peligrosos.

El uso de relaciones en el Formato XML Abierto ofrece una serie de ventajas a los desarrolladores. Las relaciones simplifican el proceso de localización de contenidos dentro del documento al no tener que interpretar el XML propio del documento para encontrar las partes, ni tampoco se necesita para localizar los enlaces a recursos internos o externos. Las relaciones permiten hacer un inventario rápido de todo el contenido del documento. Por ejemplo, si tenemos que contar el número de hojas de cálculo que tiene un libro Excel, basta con analizar las relaciones para saber cuántas partes de tipo "Hoja de cálculo" existen. Además podemos utilizar las relaciones para analizar el tipo de contenidos dentro de un documento. Esto es útil en casos donde tenemos que determinar si un documento contiene un tipo concreto de contenidos potencialmente dañino, como un objeto OLE sospechoso, o también cuando queremos, por ejemplo, extraer todas las imágenes JPEG desde un documento para utilizarlas en cualquier otro sitio.

Las relaciones también permiten a los desarrolladores la manipulación de documentos sin tener que aprender la sintaxis específica de las aplicaciones, o los tipos concretos de los contenidos. Así pues, sin un conocimiento especial de programación sobre PowerPoint, un programador podría eliminar fácilmente las diapositivas no deseadas simplemente editando las relaciones del documento.

Archivos con Macros y sin Macros

- Archivos sin macros para garantizar que ese código no se ejecutará
- Tipos de archivos con macro independientes para archivos que contienen código ejecutable.
- Se aplica a comandos VBA, Macro-Sheets de Excel y PowerPoint Action

Los documentos Office 2007 guardados con el Formato XML Abierto, por defecto, se consideran archivos sin macros y no contienen código. Esto garantiza que el código potencialmente peligroso introducido dentro de un

documento por defecto nunca se va a ejecutar de forma inesperada. Aunque los documentos pueden contener y utilizar macros en Office 2007, el usuario o el desarrollador deberán guardar explícitamente estos documentos como documentos de tipo “con macros” (*macro-enabled*). Esta medida de seguridad no afectará a la capacidad para desarrollar soluciones, y por el contrario, aumentará el nivel de seguridad dentro de las organizaciones.

Los archivos con macros tienen exactamente el mismo formato de archivo que los otros, pero contienen partes adicionales que no existen en los archivos libres de macros. Las partes adicionales dependen del tipo de automatización encontrado en el documento. Un archivo con macros que utilice VBA contendrá una parte binaria que almacena el proyecto VBA. Cuando una hoja de cálculo Excel utiliza macros de estilo Excel 4.0 (macros XML) o una presentación de PowerPoint contiene botones de acción también se guardan como archivos con macros. Si dentro de un archivo guardado como “sin macros” aparece una parte con código, ya sea introducido de forma accidental o malintencionada, las aplicaciones Office no permitirán la ejecución de ese código, bajo ningún concepto y sin excepciones.

Los desarrolladores pueden saber si hay código dentro de un documento Office 2007 antes de abrirlo. En las versiones anteriores, este “avance informativo” no era nada fácil de conseguir fuera de Office. El desarrollador puede analizar el archivo de paquete y comprobar si existen partes con código y las relaciones que mantienen antes de ejecutar Office y, con ello, el código potencialmente peligroso. Si un archivo tiene un aspecto sospechoso, el desarrollador puede eliminar cualquier parte susceptible de ejecutarse como código desde el mismo archivo, con lo que el riesgo desaparece.

Extensiones de archivo

- *Nuevas extensiones de archivo para todas las plantillas de documento*
- *Los nombres archivo sin macro, por defecto terminan con ‘x’*
- *Los nombres de archivos con macro terminan con ‘m’*

Los documentos guardados con el Formato XML Abierto en Office 2007 tendrán nuevas extensiones de archivo y permitirán a Office 2007 distinguir estos archivos de otros documentos binarios utilizados en versiones de Office

anteriores. Las nuevas extensiones toman las mismas extensiones de los archivos binarios existentes, pero les añaden una letra al final del sufijo. Las extensiones por defecto para los documentos creados en Word, Excel y PowerPoint usando el Formato XML Abierto llevarán una “x” al final: .docx, .xlsx, y .pptx, respectivamente. Otros tipos de formato de documento de Office que aprovechen el nuevo formato de archivo, como plantillas, complementos, películas de PowerPoint, etc., tendrán también extensiones nuevas.

Otro cambio introducido en Office 2007 es que existirán distintas extensiones para los archivos sin macros y con macros, a fin de distinguirlos de partida. Los documentos con macros tendrán una extensión que termina con la letra “m” en vez de con “x”. Por ejemplo, un documento de Word que contiene macros tendrá una extensión “.docm” y esto permitirá a los usuarios o paquetes de software, antes incluso de abrir el documento, saber de inmediato si puede o no contener código ejecutable.

Esta es la lista de extensiones de archivo para los tipos de documentos de Office 2007:

Tipos de archivo de Microsoft Office Word 2007		Extensión
Documento XML de Word 2007		.docx
Documento Word 2007 en formato XML con macros		.docm
Plantilla XML de Word 2007		.dotx
Plantilla de Word 2007 en formato XML con macros		.dotm

Tipos de archivo Microsoft Office Excel 2007		Extensión
Libro XML Excel 2007		.xlsx
Libro XML Excel 2007 con macros		.xlsm
Plantilla XML de Excel 2007		.xltx
Plantilla XML de Excel 2007 con macros		.xltm
Archivo binario Excel 2007		.xlsb
Complemento XML de Excel 2007 XML con macros		.xlam

Tipos de archivo Microsoft Office PowerPoint 2007		Extensión
Presentación XML de PowerPoint 2007		.pptx
Presentación XML de PowerPoint 2007 con macros		.pptm
Plantilla XML de PowerPoint 2007		.potx
Plantilla XML de PowerPoint 2007 con macros		.potm
Complemento XML de PowerPoint 2007 con macros		.ppam
Película XML de PowerPoint 2007		.ppsx
Película XML de PowerPoint 2007 con macros		.ppsm

Desarrollo de soluciones

El Formato XML Abierto de Office 2007 introduce o mejora muchos tipos de soluciones que pueden desarrollarse, relacionadas con documentos. Se puede acceder a los contenidos de un documento Office en el Formato XML Abierto utilizando cualquier herramienta o tecnología capaz de operar con archivos ZIP. El contenido puede manipularse con cualquiera de las técnicas estándar de procesamiento de XML, o para las partes que existen como formatos nativos embebidos, como imágenes, se pueden emplear herramientas adecuadas para esos tipos de objetos.

Además, al poder abrir el archivo contenedor de un documento Office 2007 manualmente (al tratarse de un archivo ZIP), surgen nuevas posibilidades interesantes para los programadores. Por ejemplo, los que desarrollan soluciones basadas en Office pueden analizar el contenido y estructura de un documento sin tener que escribir código alguno. Esta facilidad puede ser muy útil en el diseño de soluciones y para crear prototipos.

Una vez dentro de un documento Office 2007, la estructura interna permite navegar fácilmente por dentro de las partes y relaciones, ya sea para encontrar información, cambiar contenidos o eliminar elementos del documento. Mediante el empleo de XML junto con los esquemas de referencia públicos de Office, ya se pueden crear fácilmente nuevos documentos, añadir datos a los actuales o localizar contenidos concretos dentro de su cuerpo.

El resto de este Whitepaper se dedica a analizar algunos escenarios donde el Formato XML Abierto permite crear soluciones basadas en documentos. Estas son solo algunas de las casi infinitas posibilidades que se abren:

- Interoperabilidad de los datos
- Manipulación de contenidos
- Compartición y reutilización de los contenidos
- Combinación de documentos
- Seguridad de los documentos
- Gestión de información sensible
- Aplicación de formatos y estilos
- Catalogación de documentos (“perfilado”)

Interoperabilidad de los datos

La creciente popularización de XML como estándar para el intercambio de datos supone que el nuevo Formato XML Abierto permite que los datos del documento sean más fácilmente accesibles entre sistemas heterogéneos. Tanto si son los usuarios de una misma organización los que comparten los datos entre departamentos, o el intercambio se realiza entre organizaciones distintas, el uso de XML como formato por defecto para documentos de Office supone que las aplicaciones Office pueden participar de los procesos de negocio sin las limitaciones impuestas por los formatos binarios usados en versiones anteriores.

La condición de abierto del Formato XML desbloquea los datos e introduce un nuevo y amplio nivel de integración más allá de los puestos de trabajo. Por ejemplo, los desarrolladores pueden hacer referencia a una especificación publicada en el nuevo formato de archivo para crear documentos avanzados sin tener que utilizar las aplicaciones Office. Las aplicaciones de servidor podrán procesar documentos en masa para habilitar soluciones a gran escala que integren los datos corporativos con las aplicaciones Office, conocidas y utilizadas por todos los usuarios. Los protocolos estándar XML, como XPath (un lenguaje de consulta XML común) y XSLT (Extensible Stylesheet Language Transformations), pueden valer también para obtener datos desde los documentos o actualizar los contenidos dentro de un documento con datos externos.

En esta línea, podríamos diseñar un escenario para personalizar miles de documentos para su distribución a clientes. La información, extraída mediante programa desde una base de datos corporativa o una aplicación CRM se puede insertar dentro de una plantilla de documento estándar, a nivel de servidor, usando XML. La creación de estos documentos sería un proceso muy eficiente, ya que no se exige la ejecución de programas Office. Obviamente, no habría límites a la posibilidad de crear documentos Office avanzados y de alta calidad.

El uso de esquemas personalizados en Office es otra de las formas en que los documentos pueden aprovecharse para compartir datos. La información que anteriormente quedaba bloqueada dentro del formato binario ahora ya es fácilmente accesible y por tanto, los documentos pueden servir como orígenes de datos fácilmente accesibles. Los esquemas personalizados no solamente facilitan la inserción y extracción de datos, sino que añaden estructura a los documentos y permiten la validación de datos.

Manipulación de contenidos

La edición de contenidos de los documentos de Office es otro ejemplo interesante donde el Formato XML Abierto supone una mejora del proceso. La edición puede suponer actualizar pequeñas cantidades de datos, cambiar de sitio bloques enteros, eliminar partes o añadirlas. Mediante el uso de partes y relaciones, el Formato XML Abierto nos permite localizar y manipular fácilmente los contenidos. El uso de XML y los esquemas XML permiten el uso de tecnologías XML ampliamente difundidas, como XPath y XSLT, que podremos utilizar para editar datos dentro de las partes del documento prácticamente sin limitaciones.

En otras ocasiones podríamos necesitar editar el texto de la cabecera de un documento Word. Por supuesto, no tendría sentido automatizar esta actividad para un documento solo, pero ¿qué sucede en procesos de fusión empresarial o cambio de denominación de una línea de productos y la empresa quisiera actualizar centenares de documentos que contienen los nombres obsoletos? Un desarrollador podría diseñar un código que navegue por todos los documentos, localice las cabeceras en la estructura de archivo de Word y realice una consulta XPath para localizar el texto a cambiar. El nuevo texto se puede insertar, sustituir la cabecera modificada y repetir el mismo proceso para todos los documentos que necesiten actualizarse. La automatización nos puede ahorrar muchísimo tiempo y permitir un tipo de procesos que no son viables de otra manera, así como evitar los errores potenciales que pueden aparecer durante un proceso manual.

Otra situación podría ser aquella en donde un documento Office tiene que actualizarse cambiando solamente una parte de él. En un libro Excel, una hoja de cálculo con datos desactualizados o con fórmulas de cálculo no validas ya podría sustituirse con otra nueva simplemente sobrescribiendo esa parte. Este tipo de actualizaciones se aplica también a las partes binarias. Una imagen, o incluso un objeto OLE podría sustituirse directamente por otro nuevo en caso necesario. Un dibujo Visio embebido como objeto OLE dentro de documentos Office, por ejemplo, podría actualizarse sobrescribiendo la parte binaria. Los URL dentro de los vínculos pueden actualizarse para que apunten a nuevas ubicaciones.

A continuación describimos algunos escenarios más, específicos de aplicaciones Office.

Word— Manipulación de contenidos

Es una práctica frecuente insertar un texto normalizado (“boilerplate”) dentro de los documentos de Word, por ejemplo cláusulas de tipo legal, o términos de uso y condiciones que deben aparecer en todo documento de difusión pública creado por una organización. Otro ejemplo de textos normalizados son los párrafos “Acerca de” la compañía y similares,

utilizados para reforzar la imagen de marca y con fines de promoción o para anunciar noticias importantes. Word dispone de funciones como el Autotexto, capaz de insertar texto formateado, pero esta funcionalidad tiene ciertas limitaciones ya que necesita automatización de Word o la intervención manual del usuario..

Microsoft Office Word 2007 proporciona una alternativa muy flexible a los desarrolladores que permite insertar contenidos dentro de los documentos. El Formato XML Abierto permite añadir a la estructura del contenedor fragmentos de documento que y referencias que se podrán activar en todo el documento al abrirlo con Word. Esta alternativa, de gran alcance, supone que se pueden crear librerías de fragmentos de documento, obtenidas de otros documentos de Word o incluso de fuentes externas, y reutilizarse por programa cuando se necesiten, dentro de soluciones documentales basadas en Word.

Esta amplia capacidad para manipular el contenido de los documentos de Word abre una serie de escenarios muy interesantes, como el ensamblaje de documentos en aplicaciones de servidor. Volviendo al caso comentado antes, un párrafo de condiciones legales se puede insertar dentro de un documento creado en un servidor sin intervención del usuario.

Supongamos una empresa multinacional que necesita que todos sus documentos contengan un párrafo con cláusulas legales en diferentes idiomas, según el país donde se vayan a distribuir. Esta empresa puede crearse sus párrafos de texto legal en distintos idiomas como archivos .HTML y guardarlos en un servidor. Un programa de generación automática de documentos puede insertar el fragmento para el idioma adecuado como una parte guardada dentro del contenedor del documento. EL fragmento legal se reproducirá como parte del documento Word de forma natural.

Excel— Manipulación de contenidos

A fin de optimizar el rendimiento en las tareas de lectura desde disco y guardado de los archivos, así como su tamaño, los archivos, Excel 2007 solamente guardan una copia de los textos repetidos dentro del archivo Excel. Para ello Excel 2007 implementa una tabla de cadenas de texto compartidas en una parte del documento llamada [*strings.xml*]. Cada valor único de texto localizado dentro del libro se enumera una sola vez en esta parte. Las celdas individuales de las hojas de cálculo hacen referencias a esta tabla de cadenas de texto para obtener sus valores.

Así, además de optimizar el tamaño de los archivos XML de Excel, este proceso introduce ciertas oportunidades muy interesantes para crear soluciones que manipulen los contenidos. Dentro de una organización multinacional se puede aprovechar la tabla de cadenas compartidas para conseguir instancias del mismo documento en distintos idiomas. En lugar de crear libros Excel únicos para cada idioma, un mismo libro puede utilizar tablas de cadenas de texto en distintos idiomas. Otra posibilidad sería el uso de tablas de textos para realizar búsquedas por palabras clave dentro de una serie de libros Excel. El procesamiento de cadenas de texto dentro de un documento XML es mucho más rápido y sencillo que tener que manipular el modelo de objetos de Excel sobre muchos libros y hojas de cálculo.

PowerPoint— manipulación de contenidos

En una presentación de PowerPoint guardada como Formato XML Abierto, el contenido es fácilmente accesible. Al ser la primera versión de PowerPoint que ofrece un formato XML, se abren muchos nuevos escenarios que simplemente eran imposibles en versiones anteriores. Los desarrolladores ya pueden tener pleno acceso a las diapositivas y las notas como textos. Las soluciones basadas en búsqueda, creación de índices y presentación de contenidos ya son posibles. Las presentaciones gestionadas a partir de datos se pueden ya generar fácilmente usando XML. Por lo mismo, los desarrolladores pueden acceder a los archivos maestros y plantillas de presentación a partir de las partes XML para formatear mediante programa las presentaciones actuales o nuevas.

Los desarrolladores podrían aplicar diferentes técnicas para ensamblar o reutilizar contenidos de presentaciones PowerPoint creando aplicaciones que utilicen un catálogo de diapositivas almacenado independientemente de las presentaciones. Cada diapositiva se representa como una parte XML independiente, por lo que se podría optimizar la forma de almacenar y gestionar las diapositivas de PowerPoint, consideradas como datos. Se podrían incluso diseñar visores que permitiesen al usuario localizar y elegir diapositivas para crear una presentación fuera de PowerPoint. La aplicación podría, además, basarse en Web para posibilitar una gestión centralizada.

Compartir y reutilizar contenidos

La modularidad que caracteriza al Formato XML Abierto abre la posibilidad de generar contenido una sola vez y reutilizarlo en muchos documentos. Como desarrollador, puede imaginar la posibilidad de crear un número reducido de plantillas básicas, o modelos de contenidos y reutilizar sus trozos como bloques individuales para crear otros documentos.

Una tabla creada en un documento Word, por ejemplo, se podría utilizar en otros muchos documentos Word. Los diagramas, que comparten el mismo esquema entre todos los programas de Office, se pueden crear una vez y reutilizarse en muchas ocasiones en distintos tipos de documentos. La accesibilidad del formato, por sí sola, permite una variedad ilimitada de oportunidades para compartir la información.

Uno de estos escenarios podría ser aquel en que existe la intención de crear una librería de imágenes para su uso dentro de los documentos. El programador puede diseñar una solución que extraiga las imágenes de una colección de documentos Office y permitir a los usuarios reutilizarlas desde un único punto de acceso. Puesto que los documentos Office almacenan las imágenes intactas como partes binarias, la solución podría crear y mantener una librería de imágenes con suma facilidad. Después los usuarios que busquen imágenes ya utilizadas no tendrían que navegar por las carpetas y abrir todos los documentos, buscar cada una de las imágenes y copiarlas al documento que se está creando: bastará con acudir a esa aplicación que localizará las imágenes e inmediatamente quedarán disponibles para su uso en el nuevo documento.

Otra aplicación similar podría reutilizar las imágenes en miniatura obtenidas de los documentos, y darle un aspecto más visual a un sistema de gestión de documentos.

Ensamblado de documentos

Una de las funciones más solicitadas por los desarrolladores ha sido desde hace tiempo la posibilidad de crear documentos Office en un servidor sin tener que recurrir a la automatización propia de las aplicaciones Office. Las organizaciones que necesitan generar documentos complejos, con manejo avanzado de datos, o ensamblar documentos en cantidades masivas desean un proceso más eficiente para estos fines con programas Office. Técnicamente, los programas Office no han sido concebidos y no están soportados para su ejecución en un servidor.

En las ediciones de Microsoft Office 2003, la introducción de formatos de documento XML que se podían generar siguiendo los Esquemas de Referencia XML de Office 2003 ayudó a solucionar en parte esta limitación. Cualquier tecnología capaz de ensamblar XML puede crear un documento Word o Excel siempre que respete los esquemas Office. Un avance enorme en su momento que, lamentablemente, solo se puede aplicar a Excel y Word, y solo ésta última aplicación ofrece una fidelidad plena en su soporte para formato de archivo XML. Microsoft Office 2007 continúa este esfuerzo, añadiendo PowerPoint a la lista de aplicaciones

con formatos de archivo XML y garantizando la total fidelidad a los contenidos almacenados como XML en esta aplicación y en Excel, tras una revisión completa de cada una de las funcionalidades soportadas en el Formato XML Abierto.

Este avance tecnológico supone que con Office 2007 los desarrolladores pueden ya crear soluciones que generen documentos Excel, Word o PowerPoint sin siquiera tener que abrir aplicaciones Office. Simplemente han de crear XML de acuerdo con los esquemas de Office 2007 y crear los paquetes de contenidos como se indica en el Formato XML Abierto. Y aunque los esquemas Office son realmente amplios, para poder representar completamente todas las características que ofrecen los programas de Office, no todas las estructuras definidas en los formatos son imprescindibles para crear los documentos. Cada aplicación de Office es capaz de abrir el archivo con un mínimo de elementos definidos, lo que facilita la creación de la mayoría de documentos.

Debemos insistir en que el ensamblado de documentos no solamente se refiere a nuevos documentos. Por supuesto, si se siguen las reglas del Formato XML Abierto, se pueden crear documentos desde cero. Pero a menudo los procesos de ensamblado suponen crear documentos a partir de la recombinación de partes de otros documentos ya existentes, datos y otros contenidos. El nuevo Formato XML Abierto se mueve muy bien en estas circunstancias gracias a su arquitectura modular y al tratarse de contenidos basados en XML.

Un ejemplo de ensamblado de documentos es el que surge para las presentaciones de PowerPoint. En muchas organizaciones existen grandes cantidades de archivos PowerPoint susceptibles de reutilización. Pero a menudo los usuarios obtienen diapositivas de presentaciones existentes para crear otras nuevas. La localización, coordinación e integración (mediante operaciones "cortar y pegar") de diapositivas es un proceso que suele llevar tiempo, muy redundante, que en muchas organizaciones desearían poder automatizar para generar más rápidamente las presentaciones para clientes. Con Office 2007 las diapositivas individuales dentro de una presentación de PowerPoint se pueden utilizar directamente, ya que cada una está contenida dentro de su propia parte XML, en el paquete contenedor de la presentación. Una solución a medida podría aprovechar esta arquitectura para automatizar totalmente el proceso de ensamblado de presentaciones. Se podría emplear XML personalizado para mantener metadatos propios de cada diapositiva, lo que facilitaría a los usuarios la labor de búsqueda utilizando palabras clave predefinidas. Cuando el usuario selecciona una diapositiva, la solución insertaría su parte XML dentro de la presentación que se está ensamblando y generaría su correspondiente relación.

Seguridad de los documentos

La seguridad es cada vez más importante en las tecnologías de la información. El Formato XML Abierto permite a los desarrolladores un mayor nivel de seguridad al trabajar con documentos Office y crear soluciones donde la seguridad de los documentos es un aspecto fundamental. Con el Formato XML Abierto se pueden crear soluciones para identificar y eliminar vulnerabilidades conocidas o potenciales antes de que puedan causar daño.

Por ejemplo, una empresa necesita una solución para preparar documentos, bien para guardarlos en un archivo documental donde nunca debería ejecutarse código personalizado, o para enviar documentos libres de macros a clientes. Se puede crear una aplicación que elimine todo código VBA del cuerpo de los documentos Office navegando por su interior y eliminando las partes declaradas [**VBAProject.bin**] y sus correspondientes relaciones. El resultado sería un conjunto de documentos de absoluta confianza.

Lamentablemente, el código dentro de los documentos no es el único riesgo potencial de seguridad que hay que tener en cuenta. Se podrán evitar los riesgos inherentes a los objetos binarios, como objetos OLE, o incluso imágenes, localizándolos dentro de los documentos de Office y eliminando los puntos de exposición que se detecten. Por ejemplo, si un objeto OLE concreto se identifica como un peligro conocido, mediante una utilidad se podría localizar y eliminarlo, o bien someter a cuarentena todos los documentos que contengan dicho objeto. De forma similar, cualquier referencia externa introducida en un documento Office 2007 se puede identificar de forma rápida, y a partir de ahí, decidir si los recursos externos a que hace referencia son seguros o deben aplicarse medidas correctivas.

Gestión de información sensible

De la misma forma que buscamos proteger a los usuarios frente a código malintencionado, los desarrolladores pueden proteger a los usuarios del riesgo que supone el compartir los datos con destinatarios inadecuados, de forma accidental. Esta protección puede tomar la forma de *información identificable como personal* (personally identifiable information, PII) guardada dentro de un documento, o comentarios y anotaciones en el sentido de que la información marcada así no puede salir del departamento o de la compañía. Los desarrolladores podrán eliminar mediante programa ambos tipos de información directamente, sin tener que rastrear todo el documento. para eliminar los comentarios, por ejemplo, bastará con comprobar la existencia

de una relación con una parte de tipo comentario, y si existe, eliminar la parte de comentario asociada.

Aparte de la seguridad de la información privacidad y comentarios, el Formato XML Abierto permite un nivel de acceso a esta información que puede ser útil en otros casos. Se pueden crear soluciones que utilicen los datos PII para devolver una lista de documentos creados por una persona o por una organización concreta. Esta lista puede generarse sin tener que abrir Office o utilizar su modelo de objetos, gracias al Formato XML Abierto. De igual forma, una aplicación podría navegar por dentro de todos los documentos Office de una carpeta o un volumen y recopilar todos los comentarios localizados dentro de ellos. Se pueden aplicar diversos criterios para calificar los comentarios y así ayudar a los usuarios a gestionar mejor los procesos de colaboración a la hora de crear documentos.

Formato de documentos

Como sucede con otros muchos elementos de los documentos Office basados en el Formato XML Abierto, los estilos, formatos y fuentes se mantienen en partes XML independientes dentro del paquete contenedor. De nuevo, es posible crear soluciones que aprovechen esta separación. En muchas ocasiones las empresas definen métricas específicas para los documentos, y su administración y uso requieren mucho tiempo y cuidado. Sin embargo, con el nuevo formato se podrían crear aplicaciones que, por ejemplo, modifiquen o sustituyan fuentes en los documentos, sin tener que abrir Office.

Además es muy común la práctica de disponer de un documento o una serie de documentos que tienen el mismo contenido pero que se han formateado de distinta manera en cada departamento, centro de trabajo, subsidiaria, o para un cliente concreto, etc. Mediante programación se puede mantener el contenido común a toda la serie de documentos y después aplicarle los cambios de estilo en caso necesario. Para ello bastará con cambiar la parte [**styles.xml**] incluida dentro del documento Office con otra parte distinta. Esta capacidad simplifica enormemente el proceso de control de la presentación de los documentos, evitando las limitaciones derivadas del carácter confidencial o secreto de unos, o los inconvenientes y sobrecargas que impone una cantidad muy elevada de documentos a modificar, en otros casos.

Catalogación de documentos (“profiling”)

La gestión eficiente de los documentos ha sido un reto durante mucho tiempo en la práctica de las tecnologías de información. En las versiones actuales de Office, los desarrolladores tienen acceso a las típicas propiedades de los documentos, como el autor, título, asunto, etc. mediante el uso de OLE. EN el nuevo Formato XML Abierto, las propiedades del documento son accesibles directamente, ya que están guardadas en una parte independiente dentro del propio documento.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<CoreProperties
xmlns="http://schemas.microsoft.com/package/2005/06/md/core-
properties">
  <Title>Word Document Sample</Title>
  <Subject>Microsoft Word</Subject>
  <Creator>Office 2007 User</Creator>
  <Keywords/>
  <Description>Office 2007 .docx file</Description>>
  <LastModifiedBy>Office 2007 User</LastModifiedBy>
  <Revision>2</Revision>
  <DateCreated>2005-05-05T20:01:00Z</DateCreated>
  <DateModified>2005-05-05T20:02:00Z</DateModified>
</CoreProperties>
```

Parte Propiedades del Documento en archivo Word.docx (docProps\Core.xml)

Además, los documentos Office basados en el Formato XML Abierto permiten añadir nuevos datos y contenidos aparte de las propiedades específicas de Office, permitiendo de esta forma una catalogación avanzada de los documentos. Podríamos crear propiedades a medida, definidas como XML y ponerlas dentro del archivo como “otra parte” del mismo. Este XML podría luego utilizarlo una herramienta o aplicación para usos concretos.

Conclusión

Los usuarios, las organizaciones y desarrolladores. todos ellos podrán beneficiarse de las ventajas del nuevo Formato XML Abierto de Microsoft Office 2007. Al ser un formato por defecto abierto, basado en XML, el Formato XML Abierto pone a disposición de todos ellos nuevas posibilidades para crear soluciones y escenarios de muy diverso tipo. Los documentos pueden utilizarse como orígenes de datos, manipularse en ausencia de aplicaciones Office y procesarse en soluciones corporativas. Las organizaciones que combinen sus actuales inversiones en sistemas de negocio con la plataforma Microsoft Office System, Office 2007 y el nuevo formato basado en XML solo obtendrán ventajas.

Este documento analiza una versión preliminar de un producto de software que puede cambiar sustancialmente antes de la aparición de la versión comercial definitiva.

Este documento se publica con fines informativos exclusivamente y Microsoft no ofrece garantías, explícitas o implícitas, en este documento. La información contenida en este documento, incluyendo URL y otras referencias a sitios Web de Internet, está sujeta a cambios sin previo aviso. La responsabilidad con respecto a los riesgos derivados del uso o de las consecuencias del uso de este documento recae por completo en el usuario. Salvo indicaciones en sentido contrario, las compañías de ejemplo, organizaciones, productos, nombres de dominio, direcciones de correo electrónico, logos, personas, lugares y acontecimientos comentados aquí son ficticios, y ni se intenta ni debe deducirse ninguna asociación con empresas, organizaciones, productos, nombres de dominio, direcciones de correo electrónico, logos, personas, lugares o acontecimientos del mundo real.

El cumplimiento con todas las leyes aplicables sobre copyright es responsabilidad del usuario. Sin limitación de los derechos protegidos por copyright, ninguna parte de este documento puede ser reproducida, almacenada o introducida en un sistema de recuperación, o transmitida en cualquier formato o por cualquier medio (electrónico, mecánico, fotocopia, grabación o cualquier otro) para ningún propósito, sin el permiso expreso y por escrito de Microsoft Corporation.

Microsoft puede tener patentes, patentar aplicaciones, marcas comerciales, copyrights u otros derechos de propiedad intelectual cubriendo la materia analizada en este documento. Excepto que así se disponga por escrito en algún acuerdo de licencia de Microsoft, la modificación de este documento no le ofrece ninguna licencia sobre estas patentes, marcas, copyrights u otra propiedad intelectual.

© 2006 Microsoft Corporation. Todos los derechos reservados.

Microsoft, Excel, PowerPoint, Visual Basic y Windows son marcas registradas o marcas comerciales de Microsoft Corporation en Estados Unidos y/o en otros países.

Los nombres de empresas y productos mencionados en este documento pueden ser marcas de sus respectivos propietarios.