UCA Universidad de Cádiz

CITI
Área de Informática

C/Rep. Saharaui s/n, campus
Río S. Pedro, 11510 Pto. Real
Tel. 956016400
Fax 956016401
http://www.uca.es/serv/ai
area.informatica@uca.es

Formatos para el intercambio institucional de documentos en la Universidad de Cádiz

CITI, Área de Informática

En colaboración con la Oficina del Software Libre de la UCA

11 de mayo de 2006

Índice

1. Formatos de ficheros recomendados en la Universidad de Cádiz

1.1. Introducción

El lunes 11 de octubre de 2004 se publicó en el Boletín Oficial de la Universidad de Cádiz n.º 15 [1] la Normativa de Intercambio de Información Institucional en la UCA [2], por la que los documentos oficiales intercambiados o publicados digitalmente deben estar disponibles en formatos abiertos, siendo tarea del CITI [3] el mantener una lista actualizada de formatos abiertos para diferentes tipos de documentos, a los cuales deberán atenerse las publicaciones institucionales, así como difundir las instrucciones sobre cómo generarlos.

Este documento constituye una segunda versión del que fuera anteriormente publicado por el CITI, motivado por la aprobación de los nuevos estándares internacionales ISO/IEC 26300 (OpenDocument) e ISO 19005 (PDF/A).

Su objetivo es informar de cuáles son los formatos de documentos aceptables y de cómo generarlos a partir de los programas de uso más habitual hoy día en nuestra universidad. De lo primero se ocupa la sección actual, en cuyo apartado 1.4 se sintetiza el estudio sobre formatos que aparece en la sección 3. A lo segundo se dedica la sección 2, ofreciendo algunas pautas para la generación de los formatos recomendados: PDF y OpenDocument.

1.2. Terminología

Para el presente documento se empleará la siguiente catalogación

Formato cerrado o propietario Su especificación no está publicada, o lo está parcialmente, perteneciendo sus derechos a una o varias empresas que la mantienen oculta. Habitualmente se requieren técnicas de ingeniería inversa para que otras herramientas lo utilicen. Ejemplos: DOC (MS-Word), XLS (MS-Excel).

Formato abierto o estándar abierto Pertenece a alguna organización sin ánimo de lucro, con el propósito de regular el formato, que está completamente publicado. Suele existir una implementación de referencia que ayuda a desarrollar software (libre o no) totalmente compatible. Ejemplos: HTML 4.01 de W3C (ISO 15445), OpenDocument (ISO 26300), PDF/A (ISO 19005, generado a partir de la versión 1.4 del formato PDF publicado por Adobe).

Formato publicado Sus derechos pertenecen a una o varias empresas, pero su especificación está publicada totalmente y por tanto existe la posibilidad

de su uso, sin restricciones, por parte de aplicaciones no pertenecientes a la empresa propietaria. Sin embargo, el futuro del formato depende, al igual que en el caso de los formatos cerrados, de los avatares y decisiones de la empresa propietaria. Ejemplos: RTF de Microsoft, especificaciones Postscript y PDF de Adobe.

1.3. Requisitos para la selección de formatos

Atendiendo, en primer lugar, a la normativa de la UCA y, en segundo lugar, a las necesidades de los usuarios, se establecen los siguientes requisitos:

1. **Formato abierto**, por exigencia expresa de la normativa de intercambio de información institucional en la UCA.

2. **Disponibilidad de programas lectores y generadores** adecuados, valorando las siguientes características:

   a) **Libres**. La elección de los formatos debe enmarcarse dentro de la tendencia hacia el software libre de nuestra universidad, pues así no se fuerza a los destinatarios a adquirir productos comerciales, ni se fomenta la copia ilegal de software.

   b) **De buena calidad**. Los usuarios no deben verse forzados a utilizar herramientas de baja calidad, pues esto, además de los perjuicios obvios, podría generar el incumplimiento de la normativa.

   c) **Muy difundidos**. Es deseable evitar la necesidad de capacitación y adaptación a nuevas herramientas, así como generar a los destinatarios más trabajo del mínimo necesario.

   d) **Disponibles en las plataformas más usadas** en puestos de trabajo. No debe limitarse la elección de la plataforma del puesto de trabajo del usuario, al menos en el caso del personal docente. Por lo menos deben existir implementaciones adecuadas de las herramientas necesarias para las más utilizadas: MS Windows, GNU/Linux y Macintosh.

1.4. Formatos recomendados en la UCA

El CITI, por el presente documento, según mandato de la Normativa de Intercambio de Información Institucional en la UCA, publicada en el BOUCA n.º 15 el 11/10/2004, establece las siguientes conclusiones:

Se recomiendan los siguientes formatos de documentos ofimáticos:

- **PDF** para documentos que sean de lectura exclusiva para el receptor. Aunque lo realmente aconsejable el formato PDF/A (véase el apartado 3.7), ya que así se garantizaría la longevidad de los documenos, debido a la dificultad para discernir entre herramientas que generen PDF/A u otro tipo de PDF, se deja así.
- **OpenDocument** para documentos que el receptor deba, eventualmente, modificar. V. 3.3.

Se aceptan documentos de texto sencillo (V. 3.1) para situaciones donde el formato visual sea muy poco importante, como mensajes de correo electrónico o borradores, además de HTML/XHTML (V. 3.2) en versiones fijadas por el W3C [4], actualmente 4.01 y 1.0 respectivamente (por supuesto, el formato que debe ser usado en páginas web). Obsérvese que algunas aplicaciones actualmente muy extendidas, como MS-Word o MS-Frontpage, no generan HTML estándar.

Se desaconsejan aquellos formatos que, aun siendo abiertos o publicados, estén poco extendidos entre los receptores finales, como Postscript, DjVu, DVI, DocBook, TEX/LATEX, etc. (V. 3.4.)

Se rechazan todos los demás formatos propietarios, entre ellos los de MS-Office (doc, xls, ppt...) [5] y otros formatos propietarios binarios (wp de WordPerfect, etc.). También se rechazan el formato RTF (V. 3.5) y los esquemas de referencia XML de Microsoft [6].

Estas conclusiones se han obtenido como resultado del análisis realizado en la sección 3.

2. Guía para la generación de los formatos recomendados

2.1. Crear ficheros PDF usando PDFCreator

PDFCreator [7] es un programa libre para Windows [8] que crea una impresora virtual de forma que cuando se manda cualquier documento desde cualquier aplicación de Windows a «imprimir» a ella se crea un fichero en formato PDF. PDFCreator se puede descargar desde `http://sourceforge.net/projects/pdfcreator`; desde la red de la UCA también se puede instalar desde `https://cau.uca.es/software.cgi/CATEGSOFT=OFIMA`.

2.2. Crear ficheros PDF y OpenDocument usando OpenOffice.org

OpenOffice.org (abreviadamente, OOo) [9], es una *suite* ofimática libre multiplataforma capaz de abrir y guardar documentos en varios formatos, incluyendo los de MS-Office, y de exportarlos a PDF y otros formatos. Comprende una serie de programas: *Writer* es el procesador de textos, *Impress* para presentaciones, *Draw* para dibujos, *Calc* para hojas de cálculo, *Math* para fórmulas matemáticas, *Base* para base de datos y *Web* como editor de páginas web. OOo puede descargarse desde su web ([9]) o desde, por ejemplo, `ftp://ftp.rediris.es/mirror/openoffice.org`. Desde la red de la UCA también puede instalarse para Windows en `https://cau.uca.es/software.cgi/CATEGSOFT=OFIMA`. Las distribuciones normales de escritorio de sistemas GNU/Linux [10] lo traen ya incluido.

El formato nativo de OpenOffice.org desde la versión 2.0 es OpenDocument[1], por lo que si se crea un documento nuevo con este programa no hay que hacer nada especial: al archivarlo se guardará en este formato.

Si se abre en OOo un documento en formato de MS-Office (*.doc* por ejemplo), entonces a la hora de guardarlo habrá que hacerlo escogiendo del menú Archivo el elemento Guardar como… y de la lista desplegable con los formatos admitidos escoger el correspondiente OpenDocument; por ejemplo, si el documento era *.doc* ahora habrá que escoger Texto en formato OpenDocument (.odt).

Por supuesto puede convertirse un documento de OOo a PDF usando PDFCreator (V. 2.1) en Windows, como cualquier otra aplicación; pero puede hacerse más fácilmente ya que OOo permite exportar directamente el documento a PDF; esto puede hacerse de tres formas:

- mediante el menú Archivo, elemento Exportar…, escogiendo ahora del menú desplegable con los formatos admitidos el correspondiente a PDF – Portable Document Format (.pdf).
- mediante el menú Archivo, elemento Exportar en formato PDF, lo que nos lleva al punto anterior de una forma más directa.
- mediante un botón de la barra de herramientas. Dependiendo de la plataforma usada (Windows, Linux, Mac...) y la configuración de OOo, puede tener distintos aspectos, como el logotipo de Adobe (una especie de $\mathcal{A}$ roja sobre fondo blanco) o el dibujo de un documento en cuya base hay una franja roja con las letras PDF en blanco.

[1] La versión 1 no admitía OpenDocument sino un formato anterior, también basado en XML, sobre el que se construyó OpenDocument; la última versión de esta serie iba a ser la 1.1.4, pero se sacó otra nueva, la 1.1.5, con un parche para admitir el nuevo formato.

Con las dos primeras, además, se nos presentará una pantalla donde se pueden ajustar algunos detalles del fichero PDF: páginas que queremos producir, compresión y resolución de las imágenes, envío de etiquetas o notas, efectos de transición para presentaciones, formato de formularios. El botón Ayuda explica un poco estos conceptos. El tercer método (el botón) no pregunta más que el nombre del fichero destino del PDF.

3. Estudio de los formatos existentes para documentos ofimáticos

Los documentos ofimáticos son los más utilizados para el intercambio institucional, con gran diferencia; la versión actual del presente informe se centra en documentos de este tipo, aunque posteriormente pueda ser ampliado a otros tipos de documentos.

En los siguientes apartados se discuten las características de los formatos más extendidos actualmente, características que han sido utilizadas para obtener las conclusiones señaladas en la sección 1.4 y en concreto la recomendación de los formatos PDF y OpenDocument.

3.1. Texto simple

También conocido (incorrectamente) como «texto plano» o «ASCII» [12], este formato se compone de una secuencia de caracteres en alguna codificación normalizada por la organización ISO[11], como ISO-Latin1[13], ISO-Latin9 [14] o Unicode (UTF-8 [15]).

- Adecuado para: mensajes de correo electrónico.
- Ventaja: cumple con todos los requisitos enumerados en la sección 1.3
- Inconvenientes:
  - es un formato muy poco potente, no permite el resaltado del texto.
  - al no estar Unicode suficientemente extendido aún, la visualización correcta del texto depende de los códigos de caracteres empleados por emisor y receptor; aunque es fácil cambiar entre códigos, el proceso puede ser complicado y tedioso para un usuario normal.

3.2. HTML

Es un lenguaje de marcas diseñado para hipertexto [16]. Evidentemente, se excluyen de este formato extensiones no especificadas por los estándares publicados por W3C, que es el consorcio encargado de ello.

- Adecuado para: publicación en la web, formularios en páginas web, intercambio de documentos formateados sencillos.
- Ventaja: cumple con todos los requisitos enumerados en la sección 1.3
- Inconvenientes:
  - el mismo documento puede verse distinto en distintos equipos o visores (navegadores).
  - algunas herramientas de generación de HTML muy extendidas generan código HTML propietario no compatible con la especificación estándar.
  - muchas veces un documento se compone de varios ficheros, por ejemplo si tiene gráficos, por lo que resulta complicado para el usuario común el intercambio o publicación.

3.3. OpenDocument

Se trata del formato nativo de OpenOffice.org, a partir de la versión 2.0, para documentos ofimáticos[17]. Basado en XML [18], fue escogido por OASIS[2] [19] el 1 de mayo de 2005 como base para la creación de un estándar para formatos utilizados en las suites ofimáticas. El 3 de mayo de 2006 fue aprobado como un estándar ISO (ISO/IEC 26300).

A pesar de su poca difusión, esta no es tan baja como para no poder considerarse como alternativa (se estiman unos 40 millones de usuarios en el mundo, y se cree que este número aumentará), y con una política de formación adecuada que capacite a los miembros de la UCA para su empleo, es el formato más adecuado para documentos compartidos susceptibles de modificación.

Además, al ser ahora un estándar internacional, es muy probable que las administraciones públicas de Europa y buena parte del resto del mundo lo adopten como estándar ofimático obligatorio; de hecho, ya en mayo de 2004, el *Telematics between Administrations Committee (TAC)* de la Unión Europea publicó una serie de recomendaciones orientadas al fomento de un estándar como el que ahora existe.

[2] *Organization for the Advancement of Structured Information Standards*: un consorcio internacional sin ánimo de lucro que orienta el desarrollo, la convergencia y la adopción de los estándares *e-business*.

3.4. TEX/LATEX, DocBook

Lenguajes de marcado para la composición de texto y la generación de documentos en formatos adecuados para ser impresos o visualizados, como Postscript, PDF o HTML. Aunque muy extendidos en el ámbito académico, científico y de la publicación profesional, sus características los hacen poco adecuados para el uso general en la oficina.

3.5. RTF

Siglas de Rich Text Format [20], o formato de texto enriquecido, es un formato publicado propietario de Microsoft, que ha evolucionado desde la versión inicial 1.0 hasta la 1.7. Es un formato de texto, en el sentido de que el documento no se guarda en formato binario, sino que está compuesto de caracteres y órdenes de control, análogamente a TEX.

- Inconvenientes:
  - Aunque Microsoft proporciona documentación técnica sobre RTF, no existe una normativa que especifique este formato.
  - RTF no puede guardar macros, lo que también puede verse como una ventaja en cuanto a seguridad.
  - Las imágenes se guardan sin comprimir, lo que hace que los documentos con gráficos empotrados ocupen mucha memoria de almacenamiento.
  - Hay problemas de compatibilidad entre versiones: un documento RTF generado con una versión de MS-Word puede no poder abrirse desde una versión anterior.
  - A veces hay problemas con material tabular (tabuladores, filas y columnas).
  - Por último, parece que Microsoft va a abandonar este formato en favor de los esquemas de referencia de XML en las futuras versiones de MS-Office.

3.6. Esquemas de referencia XML de Microsoft

Con la versión 2003 de MS-Office, Microsoft anunció un nuevo formato basado en XML, junto con la documentación técnica de referencia y la disponibilidad de licencias gratuitas para su uso libre [21]. Esos esquemas describen cómo se almacena la información cuando los documentos se guardan como XML. Sin

entrar en detalles técnicos, básicamente en teoría cualquier fabricante de programas, ya comerciales o de software libre, podría crear procesadores de texto que entendieran estos formatos. A pesar de ello, no se va a recomendar este formato para el caso que nos ocupa debido a los siguientes

- Inconvenientes:
  - No todas las aplicaciones de MS-Office pueden hacer uso del formato XML: PowerPoint sólo guarda las presentaciones en formatos binarios propietarios, no XML.
  - Algunas características avanzadas de MS-Office 2003 solo están disponibles cuando los documentos se guardan en su formato binario propietario.
  - Hoy por hoy, este formato sólo se encuentra en la última versión de MS-Office, la 2003, y únicamente para la plataforma Windows; se incumplen pues los requisitos 2 a, c y d.
  - Microsoft no ha asegurado que vaya a publicar los esquemas de referencia XML para futuras versiones de MS-Office.
  - XML no es el formato «nativo» o predeterminado de MS-Office 2003: el usuario tiene que preocuparse de guardar el documento en este formato diciéndolo expresamente.
  - Los documentos guardados en este formato XML pueden aún contener datos binarios codificados que solamente se pueden manejar desde Windows y Office 2003.
  - Los esquemas XML no garantizan necesariamente la fidelidad de los documentos entre distintas aplicaciones de oficina.
  - Algunos aspectos legales de la licencia no están claros.

3.7. PDF

Todos los formatos enumerados hasta ahora son de tipo editable, es decir, la persona que recibe la información puede editarla y modificarla con facilidad. Sin embargo, en la mayoría de las ocasiones la información difundida está orientada a ser visualizada o impresa solamente, y no necesita ser modificada por la persona que la recibe.

PDF (Portable Document Format, o formato transportable de documentos) [22] fue creado en 1996 por la compañía Adobe Systems, Inc.[23], derivado de otro formato suyo, Postscript (PS) [24]. Tiene la propiedad de que el aspecto del documento es el mismo en pantalla que cuando se imprime en papel, y es el

mismo en cualquier plataforma. Existen lectores libres y gratuitos de muy buena calidad muy difundidos. Sobre todo el Adobe Acrobat Reader (actualmente llamado Adobe Reader), gratuito –pero no libre– y disponible para prácticamente todas las plataformas.

Diversos subconjuntos de la especificación PDF han sido estandarizados o están en proceso. Para lo que nos preocupa en este documento, conviene destacar el estándar ISO 19005-1:2005, publicado el 1 de octubre de 2005 y referido a documentos que deban preservarse en el tiempo; es decir, que se garantice que puedan leerse sin problemas en años venideros. Este estándar se conoce como PDF/A [25] y se basa en la referencia PDF 1.4 de Adobe Systems Inc., implantada en su producto Adobe Acrobat Reader 5. La principal característica es que toda la información debe estar incluida en el mismo archivo, por lo que no se admiten audio, vídeo, lanzamiento de programas, guiones Javascript, cifrado, tipos de letra externos, etc.

Otros subconjuntos de PDF estandarizados o en proceso son PDF/X (ISO 15930) para artes gráficas e impresión, PDF/E para intercambio de dibujos ingenieriles y PDF/UA para documentos accesibles universalmente.

Referencias

[1] BOUCA n.º 15: `http://www.uca.es/web/organizacion/normativa/documentos/boletines_2004/bouca_n0015.pdf`

[2] Normativa de Intercambio Institucional de Documentos en la UCA: `http://softwarelibre.uca.es/normativa_iii`

[3] Área de Informática (CITI): `http://www2.uca.es/serv/ai`

[4] Consorcio W3C: `http://www.w3c.org`

[5] Microsoft Office: `http://www.microsoft.com/spain/office`

[6] Microsoft Corp.: `http://www.microsoft.com/spain`

[7] PDFCreator: `http://www.pdfcreator.de.vu`

[8] Microsoft Windows: `http://www.microsoft.com/spain/windows`

[9] OpenOffice.org: `http://es.openoffice.org`

[10] GNU/Linux: `http://es.wikipedia.org/wiki/Linux`

[11] ISO (*International Organization for Standardization*): `http://www.iso.org`

[12] Código ASCII: http://es.wikipedia.org/wiki/ASCII

[13] Código ISO-8859-1 o ISO-Latin1: http://es.wikipedia.org/wiki/ISO-8859-1

[14] Código ISO-8859-15 o ISO-Latin9: http://en.wikipedia.org/wiki/ISO_8859-15

[15] Código UTF-8: http://es.wikipedia.org/wiki/UTF-8

[16] Lenguaje HTML: http://es.wikipedia.org/wiki/Html

[17] OpenDocument: http://es.wikipedia.org/wiki/OpenDocument

[18] Meta-lenguaje XML: http://es.wikipedia.org/wiki/XML

[19] OASIS (*Organization for the Advancement of Structured Information Standards*): http://www.oasis-open.org

[20] Formato RTF: http://es.wikipedia.org/wiki/RTF

[21] Esquemas XML de Microsoft Office 2003: http://www.microsoft.com/office/preview/itpro/fileoverview.mspx

[22] Formato PDF: http://www.adobe.com/es/products/acrobat/adobepdf.html

[23] Adobe Systems, Inc.: http://www.adobe.com/es

[24] Lenguaje PostScript: http://en.wikipedia.org/wiki/PostScript

[25] Formato PDF/A: http://en.wikipedia.org/wiki/PDF/A

[26] El Informe *Valoris* (*Comparative Assessment of Open Documents Formats Market Overview*): http://europa.eu.int/idabc/servlets/Doc?id=17982

[27] *TAC approval on conclusions and recommendations on open document formats*: http://europa.eu.int/idabc/en/document/2592/5588

[28] Política de Formatos de Documentos, Universidad de la República (Uruguay): http://www.rau.edu.uy/universidad/csdi/proyectos/0311PoliticaFormatoDocumentos_1_0_0.pdf