SECCIÓN 7

FORMATOS DE REPRESENTACIÓN DE LA INFORMACIÓN

IBERARCHIVOS IBERARQUIVOS
ARCHIVO NACIONAL COSTA RICA
UNIVERSIDAD DE COSTA RICA

MÓDULO 7

Gestión y preservación de documentos digitales

SECCIÓN 7

Formatos de representación de la información

Adaptación del Archivo Nacional de Costa Rica
Versión 1, 2024

Este curso fue traducido y adaptado por la Dirección General del Archivo Nacional de Costa Rica en colaboración con la Sección de Archivística de la Universidad de Costa Rica a partir del material original del año 2011 de la Asociación Internacional de Archivos Francófonos disponible en línea en el Portal Internacional Archivístico Francófono. Se aclara que pueden existir variaciones respecto al contenido original. Para acceder al material en francés, visite https://www.piaf-archives.org/se-former/module-7-gestion-et-archivage-des-documents-numeriques.

IBERARCHIVOS IBERARQUIVOS
ARCHIVO NACIONAL COSTA RICA
UNIVERSIDAD DE COSTA RICA

Contenido

Capítulo 1. Objetivo de la sección 5

Capítulo 2. Terminología 6

Capítulo 3. La codificación 7

- 3.1. Torre de Babel de la codificación de caracteres 8
- 3.2. Codificación universal de caracteres 11
- 3.3. ¿Dónde configurar la codificación? 11
- 3.4. Codificación numérica 12
- 3.5. Otras codificaciones 15

Capítulo 4. ¿Cuáles son los criterios para evaluar los formatos? 15

- 4.1. Apertura: formatos abierto o público 15
- 4.2. Independencia 16
- 4.3. Otros criterios por considerar 17
- 4.3. Recomendaciones existentes 17
- 4.4. Registros de formatos 19

Capítulo 5. Los formatos de datos 20
5.1. El rol especial de XML 20
5.2. Formatos de texto 23
5.3. Formatos de imagen y gráficos vectoriales 26
5.4. Formatos de audio y vídeo 28
5.5. Formatos de archivo producidos por aplicaciones «caseras» 31

Capítulo 6. ¿Está este archivo en el formato correcto? 32

Bibliografía 34

Capítulo 1. Objetivo de la sección

La representación de la información es la transformación de una o varias secuencias de bits en una información inteligible, lo cual es una premisa fundamental para poder acceder a los documentos en el corto o largo plazo.

La elección del formato de los objetos digitales constituye un punto particularmente crítico para la preservación a largo plazo.

- En el pasado, algunos documentos se perdieron porque no se tenía ningún conocimiento sobre el formato o sobre la aplicación que los creó.
- Otros tuvieron que ser sujeto de costosas migraciones, porque se habían registrado en un formato propietario, que no garantizaba su conservación. Esas migraciones han significado el estudio y desarrollo de programas informáticos que permitan releer esos datos y guardarlos en un formato neutro, independiente de todo sistema propietario.
- Peor aún, algunos documentos de oficina se volvieron a crear manualmente porque habían sido grabados en formatos propietarios totalmente cerrados, pertenecientes a una empresa que tampoco le ha dado continuidad.

El objetivo de esta sección es proporcionar al archivista un conjunto de elementos y puntos de referencia para actuar frente a este problema.

Esta sección también le ofrece al profesional de TI un enfoque diferente acerca de los programas y formatos, que no le suele ser familiar.

GLOSARIO

Formato de datos, formato de fichero o formato de representación de la información: puede definirse por el conjunto de reglas y algoritmos que permiten organizar la información en un objeto digital. Por ejemplo, el formato de datos permitirá:

- especificar la codificación de los colores de los píxeles de una imagen, definir un algoritmo de compresión de datos y la organización de estos datos (formatos PNG, TIFF...),
- especificar la organización y estructuración de la información textual a partir de la codificación elemental de caracteres (formatos SGML, XML);

En realidad, SGML y XML son en primer lugar lenguajes que contienen un conjunto de reglas, una sintaxis, palabras clave para crear documentos estructurados; cuando un documento ha sido estructurado por el lenguaje XML.

Formato digital: la representación de un objeto digital codificada como bytes, la cual define reglas sintácticas y semánticas que permiten el mapeo o correspondencia de un modelo de información a una cadena de bits y viceversa. En la mayoría de los contextos, el término formato digital es usado indistintamente con conceptos relacionados a archivos digitales como formato de archivo, envoltura de archivo, codificación de archivo, etcétera. También se le conoce como presentación digital. (Barnard, A y Voutssas, J, 2014, p. 122).

Lenguaje de marcado: un sistema de codificación legible por máquina así como sus reglas asociadas, que son utilizados para describir la estructura lógica, distribución, forma de despliegue y estilo de un cierto documentos digital. Existen varios lenguajes de marcado con diferentes reglas, propósitos y alcances: por ejemplo, el HTML, el SGML y el XML. (Barnard, A y Voutssas, J, 2014, p. 142).

Lenguaje de marcado estándar generalizado SGML: Lenguaje de marcado estándar internacional ISO 8879:1886 utilizado para la definición formal de todo tipo de documentos, de manera que los hace ser independientes del dispositivo, sistema y programa con el cual fueron realizados. (Barnard, A y Voutssas, J, 2014, p. 143).

Lenguaje de marcado extendido XML: lenguaje de marcado estándar internacional desarrollado por el World Wide Web Consortium o W3C. XML es una versión de SGML, diseñado especialmente para los documentos de la web. Permite que los diseñadores crren sus propias etiquetas, facilitando la definición, transmisión, validación e interpretación de datos entre programas y entre organizaciones. XML permite realizar funciones que no pueden lograrse con HTML. (Barnard, A y Voutssas, J, 2014, p. 143).

Capítulo 2. Terminología

La terminología del capítulo es algo técnica. Sin embargo, es absolutamente necesario ser preciso sobre los diferentes términos que se van a utilizar en esta sección.

Un formato, en su sentido amplio, es una forma estándar en la que la información se codifica para su almacenamiento en un archivo de computadora.

El formato permite definir:

- Bien el soporte físico, se hablará entonces del formato de soporte; en este caso, precisará las características físicas, por ejemplo: A4 que es un formato de papel cuyas dimensiones son 21cm de ancho x 29,7 cm de largo,
- Las características lógicas de organización de la información, entonces hablaremos de formato de datos, concepto de gran relevancia para la presente sección,
- El conjunto de características físicas y lógicas que pueden anidarse por ejemplo (VHS, Discos Compactos, cintas magnéticas), situación poco propicia para la preservación.

En esta sección se presentará una serie de ejemplos concretos sobre los formatos. El formato tendrá múltiples características, algunas de ellas son fundamentales para saber si el formato será adecuado o no desde el punto de vista de la preservación digital:

- **Formato cerrado:** un formato cerrado tiene una estructura que no es de acceso público, a excepción de aquellos que la han publicado.
- **Formato normalizado:** un formato normalizado se considerará normalizado si se ajusta a una norma emanada por una organización o ente de normalización como (ISO, AFNOR).

Atención: un estándar que describe un formato de datos, puede ser sólo un contenedor dentro del cual deben estar insertos los elementos que pueden o no ser normalizados, o ser incluso privados.

- **Formato propietario:** es un formato creado por una empresa o un propietario privado que dispone de los derechos de propiedad intelectual o del copyright correspondiente, por ejemplo, PDF, TIFF, GIF...;

Acá se pueden presentar dos escenarios:

I. El formato es propietario, pero ha sido publicado, en este caso, su propietario ha precisado los usos que están autorizados, por ejemplo, XML de Microsoft Open Office, como DOCX y XLSX.

II. El formato es propietario y no ha sido publicado (por ejemplo, los archivos producidos por Microsoft: DOC, XLS, PPT, formatos de imagen RAW y PSD de Adobe Photoshop).

- **Formato casero o formato de proyecto:** es un formato de datos definido específicamente por una aplicación casera o por un proyecto dentro de una empresa.
- **Formato público:** Se trata de un formato cuyas especificaciones fueron publicadas y son accesibles a todos sin restricción, esto no significa que el uso de este formato pueda hacerse sin restricciones.
- **Formato abierto:** Formato publicado y libre de derechos, impulsado por la comunidad, sin restricciones de uso y aplicación; es el caso de formatos definidos por el consorcio W3C (por ejemplo, HTML, PNG).

Formato estandarizado: un formato será estandarizado si se ajusta a un estándar.

Capítulo 3. La codificación

Habíamos brindado algunas generalidades sobre la codificación en la sección 3. Aquí regresamos al tema más detalladamente, ejemplificando las diferentes categorías de información que podemos codificar:

- los caracteres, en el mundo de hoy, estos son los caracteres de todas las lenguas del mundo que se deben poder codificar, es también nuestra capacidad de producir documentos multilingües con técnicas de codificación compatibles,
- los números de cualquier naturaleza y precisión. Algunas formas de codificación de números serán más adecuadas para la reducción del espacio de almacenamiento ocupado y facilidad de manipulación para los cálculos,
- la codificación de los colores sería otro ejemplo.

ATENCIÓN

No estamos obligados, para seguir con la sección, a entrar en el detalle, pero es conveniente analizar siempre el peso de los ficheros a utilizar.

3.1. Torre de Babel de la codificación de caracteres

Hasta los años 80, cada fabricante de computadoras usaba su propia codificación.

EJEMPLO

Control Data Corporation utilizaba el «Display Code», codificación de sólo 6 bits.

IBM, por su parte, había creado el código EBCDIC (Extended Binary Coded Decimal Interchange Code), un modo de codificación caracteres de 8 bits.

Esta multiplicidad de códigos daba lugar a múltiples incompatibilidades entre los distintos ordenadores.

COMPLEMENTO: CÓDIGO ASCII

El código ASCII (American Standard Code for Information Interchange) es un código de 7 bits que permite representar todos los caracteres anglosajones utilizados por una serie de fabricantes de ordenadores. También es la variante estadounidense de la norma de codificación de caracteres ISO/CEI 646.

Los códigos 0 a 31 no son caracteres visibles. Se llaman caracteres de control porque permiten realizar acciones tales como:

- Vuelta a la línea (CR significa Carriage Return),
- Pitido de sonido (BEL).

Los códigos 65 a 90 son mayúsculas.

Los códigos 97 a 122 representan las minúsculas.

	Representación gráfica	Representación binaria
Display code (Control Data)	A	000 001
ASCII 7 bits	A	100 0001
EBCDIC (IBM)	A	1100 0001

Las representaciones de la letra A en mayúsculas en Display code, ASCII y EBCDIC

Todos estos códigos han sufrido variaciones a lo largo del tiempo, pero ninguno de ellos permitía representar caracteres latinos, griegos, cirílicos, etc.

En la década de 1990, ISO estandarizó y amplió el código ASCII, creando la norma ISO 8859:

La codificación se hace sistemáticamente en 8 bits.

- Los primeros 128 caracteres son los de ASCII,
- Los siguientes 128 son específicos del idioma.

16 versiones de esta norma fueron creadas para todas las lenguas europeas, el hebreo, el cirílico, el árabe y algunos otros, siendo revisadas en 2020, manteniéndose actualmente vigentes.

COMPLEMENTO: LAS 16 VERSIONES DE LA NORMA DE CODIFICACIÓN ISO 8859

ISO 8859-1 (latín-1 o europeo occidental)	Es la parte más utilizada de ISO 8859, abarca la mayor parte de lenguas europeas occidentales: alemán, inglés, inglés vasco, catalán, danés , escocés, español, feroés y finés (parcialmente), francés (parcialmente), islandés, irlandés, el italiano, el neerlandés (parcialmente), el noruego, portugués, retrorrománticas y sueco, algunas lenguas europeas sudorientales (albanés), así como las lenguas africanas (afrikaans y swahili). El símbolo del euro y la capital, que faltaban, están en la versión revisada ISO 8859-15 (latín-9). El set de caracteres correspondientes a la ISO-8859-1 aprobado por la IANA, es la codificación que por defecto utilizaban los documentos HTML o documentos transmitidos por mensajes MIME, como las respuestas HTTP cuando el tipo de medio del documento es «text» (por ejemplo, documentos «text/html»).
ISO 8859-2 (latín-2 o europeo central)	Idiomas de Europa Central y Oriental basados en un alfabeto romano (bosnio, croata, el polaco, checo, eslovaco, esloveno y húngaro).
ISO 8859-3 (latín-3 o europeo del sur)	Para idiomas como turco (en un inicio), maltés y esperanto; suplantado por ISO 8859-9 para el turco y por unicode para el esperanto.
ISO 8859-4 (latín-4 o europeo del norte)	Estonio, letón, lituano, groenlandés y el sami.
ISO 8859-5 (cirílico)	La mayoría de las lenguas eslavas con alfabeto cirílico, incluido el bielorruso, el búlgaro, el macedonio, ruso, serbio y ucraniano (parcialmente).
ISO 8859-6 (árabe)	Cubre los caracteres más comunes del árabe. Requiere un motor de renderizado que soporte visualización bidireccional y análisis contextual.
ISO 8859-7 (griego)	La lengua griega moderna (ortografía monotónica).
ISO 8859-8 (hebreo)	El alfabeto hebreo moderno tal como se utiliza en Israel.
ISO 8859-9 (latín-5 o turco)	Similar a la ISO 8859-1, donde las letras islandesas poco se utilizan, se sustituyen por letras turcas. También se utiliza para el kurdo.

ISO 8859-10 (latín-6 o nórdico)	Reordenamiento del latín-4. Considerado más útil para las lenguas nórdicas. Las lenguas bálticas utilizan más a menudo el latín-4.
ISO 8859-11 (tailandés)	Contiene la mayoría de los glifos necesarios para la lengua tailandesa.
ISO 8859-12	Originalmente cubría los idiomas celtas. ISO 8859-12 se configuró más tarde para Latin/Devanagari, pero se abandonó en 1997, durante la 12ª reunión de ISO/IEC JTC 1/SC 2/WG 3 en Grecia. La lengua Celta se incluye en la ISO 8859-14.
ISO 8859-13 (latín-7 o báltico)	Añade algunos caracteres adicionales para las lenguas bálticas que faltaban en latín-4 y latín-6.
ISO 8859-14 (latín-8 o celta)	Cubre lenguas celtas como el irlandés (ortografía tradicional), el gaélico escocés, el manés (lengua desaparecida) y el bretón (algunas ortografías antiguas).
ISO 8859-15 (latín-9)	Revisión de 8859-1 que abandona algunos símbolos poco utilizados, sustituyéndolos con el símbolo del euro € y las letras Š, š, Ž, é, Œ, y Ÿ, lo que completa la cobertura del francés, del finés y el estonio.
ISO 8859-16 (latín-10 o europeo del sudeste)	Previsto para el albanés, el croata, el húngaro, italiano, polaco, rumano y esloveno, pero también finlandés, francés, alemán e irlandés (nueva ortografía). Esta norma apuesta más en las letras que los símbolos. El signo de moneda se sustituye por el símbolo del euro.

Las normas de codificación deben evolucionar, ya que el lenguaje evoluciona y surgen nuevas necesidades.

Ejemplo:

El símbolo € y los Œ, œœet Ÿ que faltaban para la escritura del francés se añadieron a la norma 8859-15 que constituye una revisión de la ISO latina 1 (8859-1).

Como el número de posiciones en un byte está limitado a 256, un número de símbolos poco utilizados han sido abandonados en beneficio de algunos nuevos

COMPLEMENTO: DE LA ISO LATINA-1 A LA ISO LATINA-15: DIFERENCIAS ISO 8859-15 -- ISO 8859-1

Posición	0*A4	0*A6	0*A8	0*B4	0*B8	0*BC	0*BD	0*BE
8859-1	¤	¦	¨	´	¸	¼	½	¾
8859-15	€	Š	š	Ž	ž	é	Œ	Ÿ

Por su parte, Windows utiliza la codificación Windows-1252 Idéntica a ISO 8859-1 excepto en el rango 80-9F.

Conclusión

Pero entonces, ¿cómo escribir un texto mixto francés - griego?
En general, ¿cómo codificar documentos multilingües?
¿Cómo tener en cuenta las lenguas asiáticas?
¿Cómo tener en cuenta el significado de la escritura?

3.2. Codificación universal de caracteres

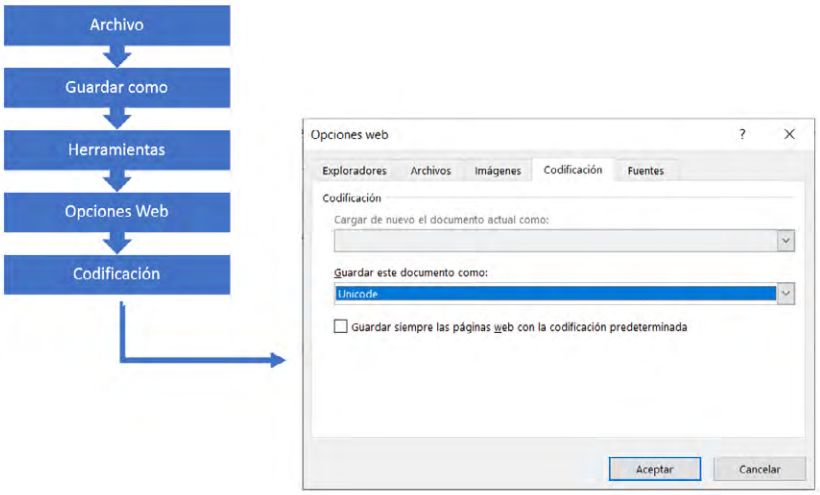
En 1991 se constituyó el Consorcio Unicode para dar respuesta a estas preguntas y tratar de solucionar definitivamente el problema de la representación de diferentes caracteres de idioma en un ordenador. Creando así un estándar universal de caracteres (unicode).

Durante los años 90s, el Consorcio Unicode y el Comité Técnico ISO/CEI JTC 1, Tecnologías de la información, a través de su subcomité SC2 (Coded character sets) unieron sus esfuerzos, ya que para esta fecha existían dos iniciativas para la codificación universal de caracteres, el estándar unicode y la norma internacional ISO 10646.

Desde esta fecha, el Consorcio Unicode e ISO han desarrollado el estándar Unicode e ISO/ IEC 10646 en conjunto. El repertorio, los nombres de los caracteres y los puntos de código de Unicode Versión 2.0 coinciden exactamente con los de ISO / IEC 10646-1: 1993 con sus primeras siete enmiendas publicadas. Después de que se publicó Unicode 3.0 en febrero de 2000, los caracteres nuevos y actualizados correspondientes ingresaron a la ISO / IEC 10646-1: 2000. En 2003, las partes 1 y 2 de ISO / IEC 10646 se combinaron en una sola parte, que desde entonces ha tenido una serie de modificaciones que agregan caracteres al estándar en sincronía aproximada con el estándar Unicode, siendo su versión más reciente la ISO/IEC 10646:2020 Information technology — Universal coded character set (UCS), por su nombre en inglés.

3.3. ¿Dónde configurar la codificación?

La forma de especificar la codificación que se desea utilizar dependerá del software. En Microsoft Word 2019 se visualiza de la siguiente forma:



Cómo especificar la codificación en Word 2019.

Se señala que, en caso de no configurar alguna especificación de codificación, Word trabajará con una codificación predeterminada de Windows, la cual es una codificación no estandarizada.

En los archivos XML de los que se hablará más adelante, se encuentra al principio de cada archivo XML la declaración, la cual indica la codificación utilizada:

?xml version=”1.0” encoding=”UTF-8”? >

3.4. Codificación numérica

Para poder representar números, se asignan valores a las posiciones de los bits.

COMPLEMENTO: CASO DE NÚMEROS ENTEROS - REPRESENTACIÓN BINARIA

Se descompone el entero en potencias de 2 y se asigna 0 o 1 a cada potencia

183 = 1x128 + 0x64 + 1x32 + 1x16 + 0x8 + 1x4 + 1x2 + 1x1

183 podrá estar representado por

10110111 o 11101101 dependiendo de la dirección que se utilizará para leer esta pequeña secuencia de bits.

En el primer caso, se pusieron los bits de peso fuerte (los asignados a las potencias superiores de 2) a izquierda,

En el otro, pusimos los bits de peso fuerte a la derecha:

- El primer enfoque se llama Big Endian (Big Endian en inglés), se utiliza en los

procesadores Intel (por ejemplo, en Windows),

- El segundo enfoque se llama pequeño-boutiste (Little Endian en inglés), que es utilizado por los Procesadores Motorola (por ejemplo, en MacOS)

Dependiendo del tamaño del número entero, se necesitan más o menos bits para su representación:

- con 8 bits, se representan números enteros positivos de 0 a 255
- con 16 bits llegaremos a 65536
- con 32 bits hasta 4 294 967 296
- con 64 bits hasta 1 844 674 073 709 551 616

De hecho, los valores son más bajos porque hay que reservar un bit para el signo.

En el caso de los números enteros - representación llamada codificada

Se puede utilizar otro método muy simple para representar un número entero:
El número 183 puede ser representado codificando sucesivamente:

- el carácter «1» de la codificación universal de caracteres (aquí idéntica a la ISO latina-1) (0110001)
- el carácter «8» de la misma codificación (0111000)
- el carácter «3» de la misma codificación (0110010)

Esta representación presenta una ventaja y una desventaja:

- la ventaja es que este valor será inmediatamente legible en cualquier editor de texto que interprete los caracteres;
- la desventaja es que para representar «183» en forma codificada, se necesitan 3 bytes, es decir, 24 bits, mientras que sólo se necesitan 8 bits para representar este número en una forma binaria; este inconveniente será marginal en muchos casos, pero será problemático cuando se trate de grandes volúmenes de datos; será más rápido y menos costoso almacenar 1 GB en forma binaria que 3 GB en forma codificada.

Por lo tanto, podemos decir que el número entero positivo 183 puede ser representado por:

- la secuencia de bits 10111 (modo binario),
- la secuencia de bits 0110001 0111000 0110010 (modo codificado)

Estas secuencias de bits sólo podrán ser interpretadas correctamente si disponemos de una información que precise el modo de representación del número, información que, a su vez, constituirá una parte de la Información de representación en el sentido del modelo OAIS.

Codificación de los números racionales y más generalmente de los números reales:

En informática, también hablaremos de números en coma flotante.

En teoría, un número real está formado por 4 elementos:

- la mantisa (número entero positivo),
- el signo del número real,
- el exponente,
- el signo del exponente.

Una vez más, podemos definir una representación llamada binaria y una representación llamada codificada. La representación codificada seguirá las mismas reglas que para los números enteros.

El número -523,12 puede codificarse con 7 caracteres: el signo «- », el separador «coma» entre la la parte entera y la parte fraccionaria y los números necesarios para la composición del número.

El número 1,6327-4 podrá ser codificado en la práctica en la forma +1,6327E-04 (el carácter «E» viene especificar que el siguiente número es un exponente. 11 bytes serán necesarios en este caso.

La representación binaria de los números reales se utilizará más a menudo para realizar cálculos. Un artificio técnico de normalización entre la mantisa y el exponente permite tener un exponente sistemáticamente positivo y, por lo tanto, no tener que almacenar el signo del exponente.

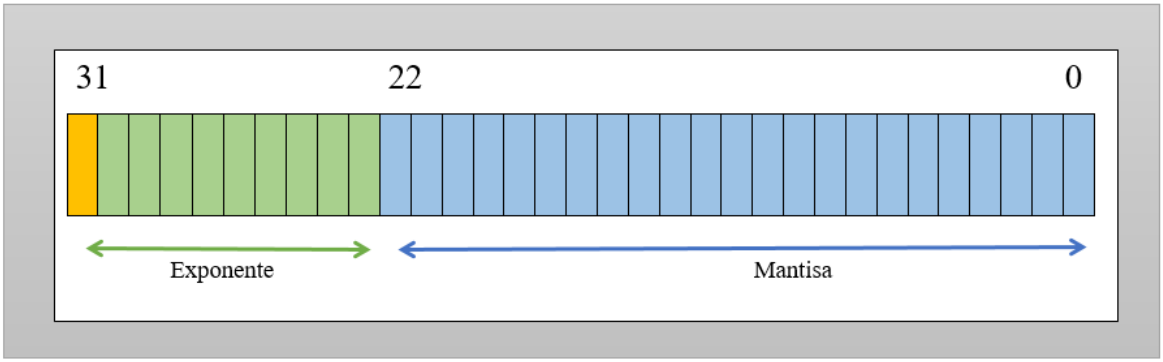
En general, se utilizarán, en función de los ordenadores y de las necesidades de precisión, representaciones de una longitud de 32, 64 o a veces 128 bits. Esto significa que la longitud acumulada de la mantisa, el exponente y el signo deben corresponder a estas longitudes.

Una vez más, hay varias convenciones diferentes en función de los sistemas operativos y

fabricantes. Uno de estos convenios está estandarizado y se recomienda su uso:

Se trata de la norma 754-2008 IEEE Standard for Floating-Point Arithmetic.

Según este estándar, un número flotante de precisión simple se almacena en 32 bits: 1 bit de signo, 8 bits para el exponente y 23 bits para la mantisa.



Representación IEEE de un número real de 32 bits. Elaboración propia a partir de PIAF.

Un número flotante de doble precisión se almacena en 64 bits:1 bit de signo, 11 bit para el exponente y 52 bits para la mantisa.

3.5. Otras codificaciones

Para la codificación de colores, se trata también de definir un estándar que permita definir el color de un píxel en una imagen, un carácter en un texto, etc.

El píxel es la unidad direccional más pequeña de la pantalla.

Existen varias posibilidades:

- Codificación de una paleta de 256 colores, en este caso, un byte bastará para definir el color,
- Codificación de 3 bytes (24 bits): 8 bits se dedican al color rojo primario, 8 bits se dedican al primario verde, 8 bits se dedican a la tonalidad primaria azul,
- Codificación de 4 bytes (32 bits): 24 bits se utilizan como en la codificación anterior, el último byte no se utiliza se utiliza para codificar, por ejemplo, información sobre transparencia.

Hay códigos específicos para el sonido, otros para el vídeo (visto como un conjunto de imágenes).

Capítulo 4. ¿Cuáles son los criterios para evaluar los formatos?

¿Cuáles son los criterios que nos ayudarán a evaluar un formato de archivo en relación con su preservación a largo plazo?
Estos criterios deberán permitir:

- identificar los formatos que facilitarán la preservación de la información,
- eliminar los formatos que planteen dificultades graves a corto o largo plazo.

Garantizar el acceso a los datos durante el mayor tiempo posible implica poder conservar la información en sus formatos originales y tener los medios para migrar a otro formato si fuera necesario. Pero cumplir con estas garantías, no es suficiente, sino que también hay que conservar las propiedades significativas que permitan representar la información.

Dada la rapidez de la evolución técnica, es muy difícil predecir la solución a adoptar. Por lo tanto, se recomienda generar los datos en formatos que respondan a estándares, con el fin de que a futuro sea mucho más sencilla la búsqueda de posibles soluciones.

A partir de este paradigma, podemos evaluar dos criterios principales, la apertura y la independencia, a los que se pueden añadir criterios complementarios.

4.1. Apertura: formatos abierto o público

Un formato público o publicado debe tener una documentación completa y accesible. Esta documentación debe estar validada y ser lo suficientemente detallada para permitir la escritura del programa, para leer los datos o convertirlos a otro formato, lo cual es una garantía esencial de seguridad para el futuro.

Es preferible que el formato sea abierto, es decir, libre de derechos, pero este punto no es determinante, sin embargo,

se recomienda comprobar las restricciones legales, ya que el uso de determinados formatos puede tener ciertas condiciones como ser de pago.

Un formato propietario ampliamente difundido (como PDF o TIFF) será preferible a un formato abierto poco utilizado.

Cuanto más se difunde un formato, más herramientas se desarrollan para explotarlo. La amplia difusión de un formato, sin embargo, no es un criterio que garantice por sí solo que ese formato sea adecuado para la preservación a largo plazo.

EJEMPLO

El formato DWG (Drawing, traducido como dibujo) es un formato cerrado utilizado por el software AutoCAD, propiedad de Autodesk.

Este formato es muy utilizado por topógrafos, geógrafos y arquitectos. Pero a pesar de su uso mundial masivo, la preservación de datos en formato DWG plantea graves dificultades.

ATENCIÓN

Tenga cuidado con los llamados formatos «contenedor» o «metarchivos» que especifican una estructura pero que permiten a su vez algoritmos de compresión: como en el caso de los formatos de imagen, los formatos de audio y vídeo.

De hecho, el formato puede ser libre de derechos, pero el algoritmo de compresión podría ser abierto y libre de derechos o no, o incluso podrían ser de paga según la elección realizada.

4.2. Independencia

Para ser adecuado para la preservación a largo plazo, un formato debe ser totalmente independiente. Esta independencia debe caracterizarse por:

- frente a otros formatos: algunos formatos pueden parecer abiertos, pero utilizan otros formatos que pueden ser cerrados o sujetos a patentes, lo cual limita el alcance de sus usos; también podrían utilizar otros elementos, como conjuntos de caracteres, que podrían ser propietarios según los casos;
- con respecto a los sistemas operativos: cuando los formatos de datos estén asociados a un sistema operativo en específico, estamos ante una restricción de acceso;

- plano económico: incluso en el caso de los formatos abiertos, los costes de desarrollo de las herramientas deben ser razonables, ello para permitir a las organizaciones o a una comunidad restringida de usuarios asumir su desarrollo;
- plano material: se trata de asegurarse de que el formato elegido no está vinculado a un dispositivo o soporte de almacenamiento específico no controlado.

4.3. Otros criterios por considerar

Cuando varias soluciones satisfacen los criterios mencionados, pueden tenerse en cuenta otros criterios como:

- la disponibilidad y el coste de las herramientas y facilidades de creación de datos, así como de las herramientas de transformación de formatos y presentación de datos,
- la posibilidad de comprobar automáticamente que un archivo se ajusta a las especificaciones del formato, y respete asimismo las normas de uso y las restricciones que se hayan definido para la preservación,
- evitar el uso de formatos innecesariamente voluminosos,
- complejidad: un formato simple es preferible a un formato completo
- la estructura del formato: cuanto más se mezclen el contenido y el estilo en el formato, será más difícil de recodificar uno sin modificar el otro o adaptar otro estilo al mismo contenido,
- la disponibilidad y el potencial de desarrollo de servicios de valor añadido, como la extracción de metadatos, o conversión a formatos de difusión.

Naturalmente, es difícil cumplir todos estos criterios complementarios. Incluso pueden ser contradictorios entre ellos: ya que un formato podría ser más simple, pero conducirá a volúmenes de datos más importantes. Por lo tanto, deben analizarse estos requisitos en función de las características y el contexto del Archivo.

ATENCIÓN

Cuanto menor sea el número de formatos de documentos aceptados y gestionados por el Archivo, menor será el riesgo de preservar la información.

4.3. Recomendaciones existentes

Actualmente existen múltiples propuestas a nivel mundial que brindan recomendaciones sobre cuáles deberían ser los formatos utilizados o las características que estos deberían de cumplir para garantizar la preservación a largo plazo, a continuación, algunos ejemplos:

- **Norma NTC-ISO 19005-1:2020 Gestión de documentos. Formato de archivo de documento electrónico para preservación a largo plazo. Parte 1: Uso del PDF 1.4 (PDF/A-1):** Esta parte de la norma NTC-ISO 19005, especifica

cómo se usa el Formato de Documento Portable (PDF) 1.4 para la preservación a largo plazo de documentos electrónicos, la cual es aplicable a los documentos que contienen combinaciones de caracteres, datos raster y vectoriales.

- **Norma UNE-ISO 15489-1:2016. Información y documentación. Gestión de documentos. Parte 1 y 2:** La cual establece un marco de preservación sistémico para documentos a través de la definición de consideraciones de orden general (Parte 1) así como una metodología de implementación (Parte 2) para un sistema de gestión de documentos.
- **Proyecto Interpares 3:** El cual busca entre otras cosas, crear modelos de evaluación capaces de medir el éxito de las soluciones de preservación que se proponga e implementen, así como emitir lineamientos orientados a los requisitos de preservación aplicables a tipos específicos de documentos de archivo.
- **UNE-ISO 23081-1:2018. Información y documentación. Procesos de gestión de documentos. Metadatos para la gestión de documentos. Parte 1:** establece un marco para la creación, gestión y uso de metadatos para la gestión de documentos, y explica los principios por los que deben regirse
- **UNE-ISO 23081-2:2021 Información y documentación. Procesos de gestión de documentos. Metadatos para la gestión de documentos. Parte 2: Elementos conceptuales y de implementación.:** Esta especificación técnica proporciona un marco para definir elementos de metadatos consistentes con los principios y las consideraciones de implementación descrita en la Norma supra citada. Presenta una aproximación lógica a los metadatos para la gestión de documentos en las organizaciones, modelos conceptuales de metadatos y una visión general del conjunto de los distintos tipos de elementos de metadatos válidos para cualquier entorno de gestión de documentos.
- **UNE-ISO/TR 15801:2019 IN Gestión de documentos. Información almacenada electrónicamente. Recomendaciones sobre confiabilidad y fiabilidad:** Define los principios que deben regirla gestión de la información considerando los siguientes aspectos: políticas de gestión de la información, deber de custodia, procedimientos y procesos, tecnologías capacitadoras y pista de auditoría.
- **Principios y requisitos funcionales para documentos en ambientes de oficina electrónica:** Publicado por el Consejo Internacional de Archivos (ICA), su propósito es definir globalmente unos principios y requisitos funcionales armonizados para el software utilizado para crear y gestionar documentos electrónicos en entornos ofimáticos
- **MoReq. Modelo de requisitos para la gestión de documentos electrónicos de archivo:** Describe un modelo de requisitos para la gestión de documentos electrónicos de archivo; se centra en los requisitos funcionales de los sistemas de gestión de documentos electrónicos de archivo.

La lista anterior no pretende ser exhaustiva, por el contrario, es una invitación a analizar con mayor detalle la multiplicidad de propuestas que existen a nivel mundial.

COMPLEMENTO

Es recomendable monitorear e identificar los formatos de archivos adecuados para el archivo digital, para lo cual puede crear una lista de formatos de archivo.

En primer lugar, se debe reconocer que tipo de contenido digital (existente o a futuro) requiere preservar, ya que eso va a definir la escogencia de los formatos. Sin embargo, no es necesario que inicie desde cero, por lo que puede ponerse en contacto con organizaciones similares.

Existen organizaciones e instituciones que han publicado su lista de formatos de archivo, por ejemplo, la Biblioteca del Congreso de los Estados Unidos publicó su lista actualizada al 2023: https://loc.gov/preservation/resources/rfs/RFS%202022-2023.pdf (material disponible en inglés).

Durante la fase de redacción de este insumo, recuerde ser realista sobre las capacidades de preservación del archivo digital y si se pueden cumplir como se requiere, a lo mejor se identificarán entre los aceptables y los preferidos.

Además, se deben tomar en cuenta revisiones posteriores que reflejan la tecnología actual y las buenas prácticas, según la investigación y análisis que realice.

4.4. Registros de formatos

La recopilación de información y documentación sobre los formatos digitales y los programas informáticos que permiten crear o leer datos organizados según estos formatos es un trabajo complejo (características, tipo, disponibilidad de la documentación, derechos de propiedad aplicables), ello ante el constante cambio y evolución de la cual son sujeto los diferentes formatos, requiriendo recursos humanos y económicos significativos.

Por lo tanto, es ilusorio imaginar que una institución pueda seguir por sí misma la evolución de los formatos digitales que gestiona.

El objetivo de los registros de formatos es trabajar de forma conjunta en la recolección, actualización e identificación de información por parte de una comunidad de usuarios.

EJEMPLO

La principal iniciativa y que está actualmente en funcionamiento:

- PRONOM en el Reino Unido (https://www.nationalarchives.gov.uk/PRONOM/)

El cual cuenta con un repositorio de investigación en GitHub que permite a los usuarios colaborar con el proyecto PRONOM y contiene guías útiles sobre cómo los usuarios pueden involucrarse en la investigación de formatos, estando próxima su actualización a la versión 109.

La información recopilada para cada formato en un registro incluye entre otras cosas:

- los nombres con los que es conocido un formato y sus variantes: por ejemplo, PDF, Adobe PDF, Portable Document Format
- Su extensión: por ejemplo, .pdf
- las especificaciones del formato
- los autores, titulares de derechos, encargados del mantenimiento
- relaciones con otros formatos derivados, versiones
- los sistemas, servicios y herramientas para la creación, lectura y validación de documentos conformes a este formato

Otra iniciativa útil en esta materia es el sitio informativo de la Biblioteca del Congreso de los Estados Unidos, este sitio ofrece un conjunto de información, publicaciones y recursos sobre los formatos y posible preservación. El sitio presenta una descripción de los formatos clasificados por tipo (texto, imagen, audio, vídeo).

Capítulo 5. Los formatos de datos

Hasta ahora se ha hablado sobre cómo representar entidades simples como números o caracteres.

Ahora se trata de representar objetos digitales más complejos que pueden contener texto, imágenes, gráficos, sonido y vídeo. Estas diferentes categorías de información también pueden recopilarse, organizadas, combinadas en documentos llamados multimedia.

La representación de estos objetos plantea otros problemas que se superponen a los anteriores.

5.1. El rol especial de XML

Origen

Desde la década de 1970 y a principios de la de 1980, IBM analiza la posibilidad de almacenar textos en ordenador. Tres de sus ingenieros diseñan un lenguaje de marcado para separar las instrucciones de estilo del contenido de documentos; Charles Goldfarb, Edward Mosher y Raymond Lorie nombran este lenguaje con sus iniciales: GML.

Los trabajos de IBM son retomados por la Organización ISO y conducen a la creación de la norma ISO 8879/1986 bautizada SGML, lo cual significa Standard Generalized Markup Language, por sus siglas en ingles.

Conceptos

El SGML introduce dos conceptos esenciales: contenido y formato separando ambos:

- Un mismo contenido puede ser utilizado con diferentes estilos

  - papel

  - pantalla

  - teléfono móvil

- el formato del documento: Hace referencias al marcado, lo cual permite asignar una verdadera semántica a los diferentes elementos de la estructura (título de documento, título del capítulo, glosario, nota a pie de página, etc).

De SGML a XML

SGML como lenguaje de marcado y descripción de documentos era tan complejo, que solo podía llevarse a cabo por especialistas (técnicos).

En 1990, con la llegada de la Web y el lenguaje HTML (Hypertext Markup Language que es un lenguaje construido sobre principios de SGML), se evidencia que SGML es muy complejo, por lo que el World Wide Web Consortium (W3C) crea un grupo de trabajo para analizar el SGML y convertirlo en una versión con mucha mayor facilidad de implementación. Esto lleva a la creación del **Extensible Markup Language** (XML) 1.0 del 10 de febrero de 1998, por parte del W3C.

Principios

XML significa en inglés Extensible Markup Language y es un lenguaje de descripción de documentos que no incluye ninguna información relativa al diseño de éstos. **XML es un metalenguaje** que permite definir otros lenguajes de marcas con objetivos diferentes. Por ese motivo se le llama 'extensible'. Por lo tanto, XML no es realmente un lenguaje en particular, sino una manera de definir lenguajes específicos que tendrán las siguientes propiedades:

- todos ellos serán conformes a XML, es decir, compartirán la misma sintaxis, las mismas reglas y podrán ser manipulados por herramientas genéricas;
- serán interoperables, es decir, se podrán crear documentos compuestos en los que cooperarán varios lenguajes XML;
- serán etiquetados de manera que evidencien la estructura del documento; el contenido y el estilo serán separados

Por lo tanto, cualquier documento XML puede ser:

- validado por un analizador sintáctico (también denominado «parser»),
- Formateado: lo cual puede ser realizado por los lenguajes CSS (Cascading Style Sheets: hojas de estilo en cascada) o XSL-FO (extensible Stylesheet Language-Formatting Objects). CSS y XSLFO son dos estándares del W3C,
- Transformado su lenguaje de estilo por ejemplo con XSLT (extensible Stylesheet Language Transformations).

Lenguajes y formatos

Cuando un documento ha sido estructurado en lenguaje XML, se conoce con anterioridad el conjunto de reglas de organización de la información dentro de este documento. Por lo tanto, XML puede ser considerado como un formato, uno abierto y legible tanto por los ordenadores y humanos.

XML utiliza el set de caracteres de Unicode y permite el uso de diferentes codificaciones incluyendo UTF-8 que es la codificación predeterminada.

XML está estandarizado por W3C.

Hasta la fecha hay cientos, tal vez miles de lenguajes XML dedicados a aplicaciones empresariales particulares. Un gran número de formatos de datos y metadatos en los más diversos campos se basan en la sintaxis XML.

Algunos son bastante genéricos para ser ampliamente utilizados:

- SMIL (Synchronized Multimedia Interfaz Language) para documentos multimedia,
- SVG (gráficos vectoriales escalables) para gráficos vectoriales en 2D,
- Xforms para formularios.

XML no es un estándar ISO pero muchas de sus normas se basan en él, muchas plataformas se benefician de este.

ATENCIÓN

Hay que añadir que, si se quiere preservar documentos XML, es necesario poder validar estos documentos al momento del ingreso al repositorio y por lo tanto disponer de las plantillas a las que hacen referencia estos archivos y preservar también estas plantillas. Puede ser una DTD (Definición de tipo de documento), esquemas XML o cualquier otro tipo de gramática estandarizada.

La adopción de modelos como DocBook, TEI (Text Encoding Initiative), textML, ALTO, es esencial en la perspectiva del archivo.

XML, un activo para la preservación de los documentos digitales

- XML es un formato abierto estandarizado por el W3 Consortium y utilizado mundialmente en todos múltiples sectores,
- XML es legible por los seres humanos y las computadoras y por lo tanto se puede convertir sin dificultad. Es sintácticamente verificable,
- XML disocia el contenido y el estilo y permite así asociar diferentes estilos a un mismo contenido,
- XML es interoperable y no depende de la plataforma informática utilizada,
- XML sólo depende de la codificación de caracteres que se realiza por UTF-8

5.2. Formatos de texto

Documentos de texto editable

También conocido como texto sin formato, texto simple (Plain Text), estos son documentos que carecen de cualquier tipo de formato tipográfico y que pueden ser leídos por cualquier ser humano.

Es un formato abierto, cuyo contenido dependerá de la codificación de caracteres, existen muchas herramientas para editar, manipular y convertir estos documentos en múltiples sistemas operativos.

EJEMPLO

- Notepad++,
- Ultraedit
- Vi
- Emac
- AkelPad

Los documentos en texto plano no plantean dificultades en cuanto a su preservación ya que se conoce la codificación utilizada. Sin embargo, son documentos “pobres” ya que contienen texto sin estructura ni estilo.

Hay una serie de formatos utilizados por las suites de oficina, los dos principales son:

- ODF (Open Document Format)
- OOXML (Office Open XML).

Formato	Descripción
ODF - Open Document Format	El Formato ODF (OpenDocumentFormato) es un formato abierto basado en un lenguaje normalizado de definición de esquema del documento RELAX NG, e incluso construido sobre el lenguaje XML. ODF fue estandarizado por el consorcio OASIS (Organization For The Avance de la información estructurada Standards) en 2005 y se convirtió en la norma ISO 26300 en 2006 siendo su última versión de 2015.

OOXML - Office Open XML	Los formatos ofrecidos por Microsoft Office como (.doc, .xls, .ppt (para las versiones más antiguas) y .docx, .xlsx y .pptx (para las versiones recientes 2007 en adelante) fueron formatos cerrados, y no fueron publicados sino hasta principios de los años 2000. Estos formatos, así como el paquete de software de Microsoft Office han evolucionado a razón de liberar una nueva versión cada dos años desde 1990. **Atención** **¡Muy importante!** **La retro compatibilidad, es decir la capacidad de leer un archivo creado con una versión anterior por parte de una versión nueva en la suite Office, no está asegurada después de 10 años.** Microsoft ha desarrollado su propio formato abierto de archivo el MS-OOXML el cual está basado en XML para documentos de oficina. Abarca los documentos del procesador de texto, hojas de cálculo, presentaciones, gráficos, diagramas, figuras. MS-OOXML fue adoptado por primera vez como estándar en 2006 por parte del ECMA, una organización privada de estándares internacionales supuestamente como un estándar abierto. En 2008, la Organización Internacional de Normalización (ISO) también aprobó el MS-OOXML, como Estándar Abierto internacional bajo la Norma ISO/IEC 29500.

La situación en relación con los formatos de los documentos de oficina ha cambiado drásticamente en los últimos años.

ODF parece ser una de las opciones para el archivo a largo plazo de documentos editables, pero no podemos predecir lo que sucederá a futuro.

La realidad muestra que se debe ser cauteloso en este aspecto.

Documentos de texto no editable

El formato PDF (Portable Document Format) y su versión PDF/A dedicada al archivo

Formato	Descripción
PDF 1.7 y 2.0	PDF es un formato propietario publicado. Pertenece a la empresa Adobe. Es un formato contenedor. Este permite contener otros tipos de formato de datos como imágenes, sonido, vídeo, etc. Hay muchas herramientas para manipular este formato. La política de Adobe es distribuir gratis las herramientas de lectura y vender las herramientas de creación. El hecho de que sea un formato contenedor obliga a comprobar rigurosamente que los archivos PDF destinados a la preservación estén construidos únicamente con los elementos previstos. El formato puede incluir metadatos en formato XMP (Extensible Metadata Platform). En julio de 2008, la versión 1.7 de PDF se convirtió en ISO 32000-1 2008 (Gestión de documentos - Formato de documento portátil - Parte 1: PDF 1.7), siendo su última revisión en 2018. A su vez en 2020 la versión 2.0 se convirtió en ISO 32000-2:2020 (Gestión de Documentos- Formato de Documento Portátil- Parte 2: PDF 2.0). A pesar de ser ambos formatos normalizados ello no exime la necesidad de tomar las precauciones indispensables ya explicadas.
PDF/A	La versión 1.4 de PDF fue la base sobre la que se definió en 2005, la norma ISO 19005-1: Gestión de documentos. Formato de fichero de documento electrónico para la conservación a largo plazo. Parte 1: Uso del PDF 1.4 (PDF/A-1). Revisada en 2020, PDF/A contiene una serie de restricciones en comparación con PDF pero integra todos los elementos necesarios para el acceso del documento y, en particular, las propiedades de visualización. En el año 2011 se publicó la especificación PDF /A -2 (Formato de fichero de documento electrónico para la conservación a largo plazo. Parte 2), y está construida alrededor de la versión 1.7 del formato PDF y por último A finales del año 2012 se publicó la tercera especificación el PDF /A-3, que soporta la incrustación de archivos en otros formatos. Normalmente al utilizar estos estándares el resultado es un aumento del volumen de un archivo, pero en contrapartida, una independencia de estos ficheros en relación con las plataformas en cuáles se utilizan.

ATENCIÓN

El formato PDF está destinado a los documentos no editables. Presenta la ventaja de poder representar la presentación original del documento de manera fiel, garantía que no siempre puede ser totalmente asegurada por los formatos ofimáticos editables. Sin embargo, no hay que creer que un documento PDF no pueda ser modificado de manera mal intencionada.

5.3. Formatos de imagen y gráficos vectoriales

Hay dos tipos de formatos

- Las imágenes de píxel. Una imagen en mapa de bits, imagen ráster o imagen de pixeles o píxeles, es una estructura o fichero de datos que representa una rejilla rectangular de píxeles o puntos de color, denominada matriz, que se puede visualizar en un monitor, papel u otro dispositivo de representación,
- Las imágenes vectoriales. Una imagen vectorial o gráfica es un tipo de imagen definida en un plano, conectada por líneas y curvas, dando como resultado formas basadas en ecuaciones matemáticas. Y debido a esto, si acercas o alejas el zoom, las líneas, curvas o puntos siempre permanecen suaves.

Formatos de imagen con descripción de píxeles

Abreviación	Nombre y estado	Características
GIF	• Formato de gráficos intercambiable • Formato propietario publicado • Limitaciones relacionadas con su patente	Posee una patente de propiedad la cual podría significar algunas restricciones en su uso.
TIFF	• Formato de archivo de imagen etiquetado • Formato publicado • Propiedad de la empresa Adobe • Sin licencia de uso	Es un formato contenedor: Define una estructura. Permite incluir los perfiles ICC (International Color Consortium) en el archivo. El perfil ICC permite la gestión del color independiente de las plataformas y dispositivos. La imagen podrá registrarse según diferentes algoritmos de compresión según la elección del usuario. Las imágenes también pueden ser guardadas sin compresión.

		El éxito de este formato se debe a dos razones principales: • Permite registrar las imágenes en blanco y negro con el algoritmo ITU T6 que había sido creado para la transmisión por medio del fax: este algoritmo es muy eficaz y ofrece tasas de relación de compresión elevadas sin pérdidas. • Este formato también ofrece la posibilidad de almacenar metadatos técnicos muy complejos. TIFF se utiliza como formato de escaneo.
JPEG y JPEG2000	• JPEG (Joint Photographic Expert Group) es un formato publicado y abierto. • JPEG publicado como Norma ISO/IEC IS 10918-1, su más reciente versión ISO/IEC 10918-1:1994/Cor.1:2005 • JPEG2000 es un formato publicado y abierto publicado como Norma ISO/CEI 15444-1, en su más reciente revisión del 2019.	La norma JPEG describe el algoritmo de compresión y la información mínima para usarlo. La razón de su éxito es que ha sido implementado de forma masiva como código abierto, además de ser admitido por todos los navegadores de Internet. Su principal inconveniente es que utiliza un algoritmo de compresión con pérdida (aunque esta tasa de compresión puede ser muy leve). JPEG no es propiamente un formato de fichero, JPEG se basa en el formato JFIF(formato de intercambio de ficheros JPEG), el cual define especificaciones complementarias para el formato contenedor que contiene los datos de imagen codificados con el algoritmo JPEG Con JPEG2000, la compresión es una de las mejoras importantes de este formato, en comparación con el JPEG, aunque la calidad de es similar igual.
JBIG	• Formato trabajado por el Joint Bi-level Image experts Group, el cual desarrollo la Norma ISO/IEC IS 11544, la cual esta vigente y cuenta con una última revisión en 2021. • Formato publicado abierto. • Limitación de las patentes en el algoritmo de compresión	JBIG utiliza un algoritmo de compresión sin pérdida. Puede también ser utilizado para la codificación a niveles de gris e imágenes a color con un número limitado de bits por píxel. El inconveniente es su algoritmo de compresión, el cual está sujeto a una patente propiedad de IBM, Mitsubishi y Lucent. Es probablemente una de las razones de su escasa difusión.

PNG Portable network Graphics	• Formato publicado abierto • Estándar del W3C • Norma ISO/IEC 15948:2004 revisado y validado en 2021.	Formato que soporta imágenes en escala de grises o en color. Es aceptado por la mayoría de los navegadores modernos.

Formatos con descripción vectorial

Abreviación	Nombre y estatus	Principales características
SVG	• SVG Vector Scalable Graphics • Estándar abierto del W3C	Familia de formatos basados en lenguaje XML abiertamente documentados y utilizados para generar gráficos vectoriales bidimensionales para su uso en la web, son desarrollados por el World Wide Web Consortium (W3C).
DWG	• Abreviación de DraWinG (Dibujo) • Formato cerrado, propiedad de la sociedad Autodesk, distribuidor del software Autocad	DWG es el formato de archivo nativo patentado para AutoCAD, una de las herramientas de diseño por computadora más populares. DWG es un formato binario compacto que almacena y describe el contenido de metadatos y datos de diseño 2D y 3D.

5.4. Formatos de audio y vídeo

Los formatos audiovisuales y sonoros constituyen un ámbito técnicamente complejo. Sin embargo, los criterios de evaluación de los formatos en relación con la preservación digital se aplican en su totalidad.

Se hará una distinción sistemática entre la especificación del formato que define el modo de encapsulación del contenido, es decir, la organización de la información dentro de un contenedor y la codificación propiamente dicha de la información que utilizará, a menudo, un algoritmo de compresión de datos para reducir su volumen.

Daremos algunas indicaciones generales sobre una selección de formatos existentes.

ATENCIÓN

Es frecuente separar los formatos de preservación, es decir, los que permiten conservar toda la información y los formatos de difusión, aquellos que proporcionan en respuesta a una solicitud por parte del consumidor.

Formatos de audio

Los formatos de audio son formatos «envolventes», llamados también formatos «contenedor». Dentro de estos formatos, la información de audio se puede codificar y comprimir de varias maneras, esto con el fin de definir claramente como representar el audio, por lo que es conveniente especificar siempre el tipo de codificación que se utilizará.

Hay recomendaciones de formatos sonoros emitidas por la IASA (International Association of Sound and Audiovisual Archives). Lamentablemente, no hay equivalente para el vídeo.

Se presenta una breve síntesis de los principales formatos.

Abreviación	Nombres y estatus	Principales características
Wave	• Formato publicado, propiedad de Microsoft.	Formato de archivo para audio, de tipo contenedor que puede incorporar un flujo de bits de audio con otros fragmentos de datos. Una codificación de flujo de bits común es LPCM (modulación de código de pulso lineal). Los proyectos de preservación que emplean el reformateo generalmente usan una de las variantes de Broadcast WAVE. Completamente documentado. Formato patentado desarrollado por Microsoft e IBM como parte del formato de archivo de intercambio de recursos (RIFF) para Windows 3.1, con documentación disponible gratuitamente.
AIFF	● Audio Interchange File Format ● Formato publicado, propiedad de Apple	Formato de archivo para sonido que envuelve varios flujos de bits de sonido, desde forma de onda sin comprimir hasta MIDI.
MP3	● MPEG Audio Layer 3 ● Algoritmo de compresión (ISO/CEI IS 11172-3 y ISO/CEI IS 13818-3).	Formato de audio dotado de un algoritmo de compresión capaz de reducir gran cantidad de datos necesaria para restituir audio con pérdida de calidad sonora aceptable para el oído humano. MP3 es el nombre, para la codificación de audio MPEG Layer III se define en dos familias de especificaciones ISO/IEC (MPEG-1: 11172-3 y MPEG-2: 13818-3), siendo un estándar abierto. No posee restricciones para la estructura del formato

OGG	• Formato publicado y abierto Formato abierto promovido por la fundación Xiph.org.	Este formato debe utilizarse con la codificación Vorbis, también definida por esta fundación. Vorbis es un algoritmo de compresión y descompresión de audio digital, abierto y libre, de mejor rendimiento en términos de calidad y tasa de compresión que el formato MP3.

Formatos de vídeo

Las características de un formato de video son complejas por lo que la elección de un formato para la preservación requiere de un análisis específico.

Se proponen algunos elementos como base de reflexión introductoria y no exhaustiva.

Abreviación	Nombre y estatus	Principales características
MPEG	• Moving Pictures Expert Group • Formato abierto • Conjunto de normas ISO	Este formato corresponde a una familia de normas ISO: • MPEG-1: las primeras películas en Internet (ISO/CEI 11172-1 a 5) • MPEG-2: la televisión digital actual • MPEG - 4: La televisión Digital Terrestre • MPEG-7, MPEG-21: futuros normas de composición de escenas, muy ricas en metadatos. La más utilizada actualmente es MPEG-4 que requiere ser utilizada con una codificación llamada H264. H.264, o MPEG-4 AVC (Advanced Video Coding), es una norma de codificación de vídeo desarrollada conjuntamente por la UIT (Unión Internacional de Telecomunicaciones) y la ISO. La norma UIT-T H.264 y la norma MPEG-4 , Parte 10 (ISO/CEI 14496-10) son técnicamente idénticas.
MJPEG 2000	• Motion JPEG 2000 • Formato publicado y abierto	Estándar abierto. Desarrollado conjuntamente por el Grupo de expertos en imágenes en movimiento (MPEG), un grupo de trabajo dentro de la estructura del comité ISO/IEC conocido formalmente como ISO/IEC JTC 1/SC 29. Los cuadros MJP2 se representan como entidades separadas codificadas con J2K_C (con pérdida o sin pérdida).

OGG	• Formato publicado y abierto	Formato abierto promovido por la fundación Xiph.org. Este formato debe utilizarse con la codificación Theora. Este es el modo de compresión de vídeo libre y sin patentes promovidas por la misma fundación.
Matroska	• (Матрёшка o Muñeca rusa) • Formato abierto	Matroska es un formato que puede agrupar en un mismo archivo varias pistas de vídeo, audio, subtítulos y capítulos. Como en MPEG 4, se recomienda la codificación H264.
AVI	• Audio Video Interleave • Formato propietario (Microsoft) • Formato contenedor publicado	Este formato contenedor permite cualquier codificación. En modo no comprimido, los archivos son muy pesados.

5.5. Formatos de archivo producidos por aplicaciones «caseras»

Hasta ahora se han mencionado principalmente de formatos abiertos o propietarios, que están disponibles en el mercado y para los que existen programas informáticos de escritura, lectura, conversión, etc.

Muchas empresas o instituciones desarrollan sus propias aplicaciones de software para satisfacer sus necesidades. Los datos digitales producidos por estas aplicaciones tienen, por tanto, un formato propio.

Al no basarse en documentación descriptiva disponible o aprobada por un organismo de normalización, o por parte de su propietario, será necesario aquí, que el desarrollador elabore su propia documentación descriptiva del formato. Esta descripción deberá imperativamente:

- ser completa,
- ser precisa,
- haber sido validada rigurosamente.

ATENCIÓN

La validez y completitud de esta descripción del formato de aplicación «casera» son elementos determinantes para la preservación de los datos obtenidos de esta aplicación. Cualquier incumplimiento de esta documentación conlleva inmediatamente un riesgo importante de pérdida o interpretación errónea de la información preservada.

El detalle de la forma en que debe presentarse esta documentación sale del alcance de este curso, pero es conveniente saber:

- Existen métodos de descripción formal. Estos métodos permiten presentar una descripción del formato que será interpretado tanto por personas como por programas informáticos,
- Las herramientas permiten garantizar la coherencia entre un documento digital y la descripción formal de su formato.

Capítulo 6. ¿Está este archivo en el formato correcto?

La extensión del nombre de fichero (.pdf, .doc, .xml...) permite conocer rápidamente el tipo de datos, sin embargo, la extensión, (especialmente en ambiente windows) es insuficiente, ya que varios tipos de datos comparten la misma extensión.

Por ejemplo, para PDF podemos tener:

- Systems Management Server (SMS) Package Description File (Microsoft Corporation)
- ArcView Preferences Definition File (ESRI)
- Netware Printer Definition File
- **Acrobat Portable Document Format (Adobe Systems Inc.)**
- P-CAD Database Interchange Format (Altium Limited)
- Package Definition File
- etc.

COMPLEMENTO

«Magic number» o número mágico es un número que, situado al comienzo de un fichero, indica su tipo a los programas que lo tratan. Esta técnica es utilizada por diferentes sistemas operativos como MacOS y Unix. Magic number ofrece mayor fiabilidad a la que brinda la extensión cuyo valor puede cambiarse voluntariamente o muy fácilmente.

Por ejemplo, el número mágico para los archivos PDF tendrá como valor %PDF-1.

Para los ficheros TIFF tendrá para magic number MM. * o II* según como esté constituido el fichero.

Sin embargo, sólo el análisis completo de los formatos es lo que nos permitirá asegurarnos de que el archivo se ajusta a las especificaciones del formato al que se refiere.

ATENCIÓN

La validación de los formatos de los ficheros que ingresan a un repositorio constituye una operación crítica. Si los ficheros transferidos por el productor no cumplen plenamente las especificaciones de formato y, además, si las no conformidades no se identifican explícitamente en las herramientas de lectura actuales, se genera un riesgo importante para el repositorio, que se hace responsable de preservar la información una vez que este es transferido y aceptado.

Por ejemplo, es común que los archivos PDF no cumplan la especificación publicada por Adobe sin que las anomalías sean señaladas por la herramienta de lectura gratuita Acrobat Reader. La validación de los formatos de entrada implica el uso y, en su caso, el desarrollo de herramientas de control de estos formatos y la aplicación de procedimientos muy rigurosos.

EJEMPLO

A continuación, se presentan algunos ejemplos que evidencian las dificultades que presenta, el no validar los documentos por parte de los repositorios documentales.

- Los ficheros HTML, pueden contener errores de sintaxis que deberían ser rechazados por la herramienta de validación (repositorios), por el contrario, si se aceptan, algunos de estos errores pueden dificultar la migración del formato a futuro.
- La validación de archivos XML implica referirse a esquemas o DTD externas. Esto exige, por lo tanto, haber previamente recuperado estos esquemas o modelos y conservarlos en el sistema de preservación,
- la validación de archivos de vídeo en formato MPEG es compleja debido a la gran permisividad de estructura de este formato,
- No existe a la fecha una solución para la conversión de archivos gráficos (tipo Autocad).

Se han desarrollado varios programas de validación de formatos.

EJEMPLO

- JHOVE (JSTOR/Harvard Object Validation Environment) que permite validar la conformidad de los ficheros en relación con varios formatos, como AIFF (Audio Interchange File Format, formato de audio de Apple), GIF, HTML, PDF, TIFF, JPEG (Joint Photographic Experts Group), XML, WAVE (formato de audio de Microsoft), etc. JHOVE es un software libre bajo licencia GPL de GNU,
- DROID (Digital Record Object Identification) es una herramienta de código abierto proporcionada por el Archivo Nacional del Reino Unido. Se basa en el registro de formato PRONOM y vincula la identificación del formato con los documentos técnicos correspondientes disponibles en el registro.

Otra iniciativa útil es la del CINES, que ha puesto en línea un servicio de validación de formatos utilizando como base JHOVE, DROID y otras herramientas. Este servicio, llamado «Facile», permite no tener que instalar ningún software de validación de formato.

ATENCIÓN

La validación de la información digital es la clave de la preservación a largo plazo, ya que garantiza el acceso: datos y metadatos.

Pero también puede llegar a ser la fuente de múltiples dificultades técnicas, de ahí la necesidad del trabajo multidisciplinario, el compartir las experiencias y herramientas.

Bibliografía

Asociación Española de Normalización y Certificación. (2016). Norma Internacional ISO 15489-1:2016 Información y documentación. Gestión de documentos. Parte 1: Conceptos y principios. España: AENOR.

BANAT-BERGER F., HUC C., DUPLOUY L., *L'Archivage numérique à long terme, les débuts de la maturité?* (Primera obra de síntesis sobre el archivo digital en lengua francesa) Paris, La Documentation française, 2009

BANAT-BERGER F., HUC C., Module 7 - Gestion et archivage des documents numériques. Portail International Archivistique Francophone. 2011. https://www.piaf-archives.org/se-former/module-7-gestion-et-archivage-des-documents-numeriques (Se identifica en el texto como PIAF)

Barnard, A y Voutssas, J (2014). Glosario de Preservación Archivística Digital Versión 4.0. Universidad Nacional Autónoma de México. https://iibi.unam.mx/archivistica/glosario_preservacion_archivistica_digital_v4.0.pdf

Borrell Saburit, A., Cueto González, A. E., Marteautus Medina, Y., Navarrete, C., & Mazorra Fernández, Y. (2007). Selección de sitios y portales especializados en conservación y restauración de documentos. Acimed, 15(3), 0-0.

Consejo Internacional de Archivos. (2013). Proyecto Interpares. *Desarrollo de Políticas y Procedimientos para la Preservación Digital*. Disponible en: http://interpares.org/ip3/display_file.cfm?doc=ip3_canada_gs12_module_2_sp.pdf

Consejo Internacional de Archivos. (2013). Proyecto Interpares. Desarrollo de Políticas y Procedimientos para la Preservación Digital. Disponible en:

http://interpares.org/ip3/display_file.cfm?doc=ip3_canada_gs12_module_1_sp.pdf

Council on Library and Information Resources (S.F). Management of Digital Information: A File Format Investigation., Disponible en: https://www.clir.org/pubs/reports/pub93/.

José Luis Alonso Berrocal Alonso, Zazo Rodriguez, Figuerola Carlos (2000). *SGML/XML: Desarrollo en entornos documentales.* Disponible en: http://angelzazo.usal.es/data/_uploaded/file/alonso2000sgml.pdf

Kenney, Anne R. (2010) Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems (Tutorial de la Cornell University Library). Última actualización mayo 2010. Disponible en Internet: http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html

Organización Internacional de Normalización. (1998). ISO 8859-1:1987 Information processing — 8-bit single-byte coded graphic character sets — Part 1: Latin alphabet No. 1

Organización Internacional de Normalización. (1999). ISO 8859-2:1987 Information processing — 8-bit single byte coded graphic character sets — Part 2: Latin alphabet No. 2

Organización Internacional de Normalización. (1999). ISO 8859-3:1988 Information processing — 8-bit single-byte coded graphic character sets — Part 3: Latin alphabet No. 3

Organización Internacional de Normalización. (1998). ISO 8859-4:1988 Information processing — 8-bit single-byte coded graphic character sets — Part 4: Latin alphabet No. 4

Organización Internacional de Normalización. (2020). ISO/IEC 8859-5:1999 Information technology — 8-bit single-byte coded graphic character sets — Part 5: Latin/Cyrillic alphabet

Organización Internacional de Normalización. (1999). ISO 8859-6:1987 Information processing — 8-Bit single-byte coded graphic character sets — Part 6: Latin/Arabic alphabet

Organización Internacional de Normalización. (2003). ISO 8859-7:1987 Information processing — 8-bit single-byte coded graphic character sets — Part 7: Latin/Greek alphabet

Organización Internacional de Normalización. (1999). ISO 8859-8:1988 Information processing — 8-bit single-byte coded graphic character sets — Part 8: Latin/Hebrew alphabet

Organización Internacional de Normalización. (2020). ISO/IEC 8859-9:1999 Information technology — 8-bit single-byte coded graphic character sets — Part 9: Latin alphabet No. 5

Organización Internacional de Normalización. (2020). ISO/IEC 8859-10:1998 Information technology — 8-bit single-byte coded graphic character sets — Part 10: Latin alphabet No. 6

Organización Internacional de Normalización. (2020). ISO/IEC 8859-11:2001 Information technology — 8-bit single-byte coded graphic character sets — Part 11: Latin/Thai alphabet

Organización Internacional de Normalización. (2020). ISO/IEC 8859-13:1998 Information technology — 8-bit single-byte coded graphic character sets — Part 13: Latin alphabet No. 7

Organización Internacional de Normalización. (2020). ISO/IEC 8859-14:1998 Information technology — 8-bit single-byte coded graphic character sets — Part 14: Latin alphabet No. 8 (Celtic)

Organización Internacional de Normalización. (2020). ISO/IEC 8859-15:1999 Information technology — 8-bit single-byte coded graphic character sets — Part 15: Latin alphabet No. 9

Organización Internacional de Normalización. (2020). ISO/IEC 8859-16:2001 Information technology — 8-bit single-byte coded graphic character sets — Part 16: Latin alphabet No. 10

Organización Internacional de Normalización. (2020). *Norma NTC-ISO 19005-1:2020 Gestión de documentos. Formato de archivo de documento electrónico para preservación a largo plazo. Parte 1: Uso del PDF 1.4 (PDF/A-1)*.

Organización Internacional de Normalización. (2016). Norma UNE-ISO 15489-1:2016. Información y documentación. Gestión de documentos. Parte 1 y 2.

Organización Internacional de Normalización. (2018). UNE-ISO 23081-1:2018. Información y documentación. Procesos de gestión de documentos. Metadatos para la gestión de documentos. Parte 1

Organización Internacional de Normalización. (2021). UNE-ISO 23081-2:2021 Información y documentación. Procesos de gestión de documentos. Metadatos para la gestión de documentos. Parte 2: Elementos conceptuales y de implementación

Organización Internacional de Normalización. (2019) UNE-ISO/TR 15801:2019 IN Gestión de documentos. Información almacenada electrónicamente. Recomendaciones sobre confiabilidad y fiabilidad

Unión Europea. (2010). *Modelo de requisitos para la gestión de documentos electrónicos de archivo*. Disponible en: https://www.moreq.info/specification

IBERARCHIVOS
IBERARQUIVOS

ARCHIVO NACIONAL
COSTA RICA

UNIVERSIDAD DE
COSTA RICA