SEÇÃO 7

FORMATOS DE REPRESENTAÇÃO DA INFORMAÇÃO

IBERARCHIVOS IBERARQUIVOS
ARCHIVO NACIONAL COSTA RICA
UNIVERSIDAD DE COSTA RICA

MÓDULO 7

Gestão e preservação de documentos digitais

SEÇÃO 7

Formatos de representação da informação

Adaptação do Arquivo Nacional da Costa Rica
Versão 1, 2024

Este curso foi traduzido e adaptado pela Direção Geral do Arquivo Nacional da Costa Rica, em colaboração com a Seção de Arquivologia da Universidade da Costa Rica, a partir do material original de 2011 da Associação Internacional de Arquivos Francófonos, disponível online no Portal Internacional Arquivístico Francófono. Esclarece-se que podem existir variações em relação ao conteúdo original. Para acessar o material em francês, 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

Conteúdo

Capítulo 1. Objetivo da seção

A representação da informação é a transformação de uma ou várias sequências de bits em uma informação inteligível, o que é uma premissa fundamental para poder acessar os documentos a curto ou longo prazo.

A escolha do formato dos objetos digitais constitui um ponto particularmente crítico para a preservação a longo prazo.

- No passado, alguns documentos foram perdidos porque não havia nenhum conhecimento sobre o formato ou sobre o aplicativo que os criou.
- Outros tiveram que passar por migrações custosas, porque foram registrados em um formato proprietário, que não garantia sua preservação. Essas migrações significaram o estudo e o desenvolvimento de programas de computador que permitissem reler esses dados e salvá-los em um formato neutro, independente de qualquer sistema proprietário.
- Pior ainda, alguns documentos de escritório tiveram que ser recriados manualmente porque haviam sido gravados em formatos proprietários totalmente fechados, pertencentes a uma empresa que também não lhes deu continuidade.

O objetivo desta seção é fornecer ao arquivista um conjunto de elementos e pontos de referência para agir diante desse problema.

Esta seção também oferece ao profissional de TI uma abordagem diferente sobre os programas e formatos, que normalmente não lhe é familiar.

GLOSSÁRIO

Formato de dados, formato de arquivo ou formato de representação da informação: pode ser definido pelo conjunto de regras e algoritmos que permitem organizar a informação em um objeto digital. Por exemplo, o formato de dados permitirá:

- especificar a codificação das cores dos pixels de uma imagem, definir um algoritmo de compressão de dados e a organização desses dados (formatos PNG, TIFF...),
- especificar a organização e estruturação da informação textual a partir da codificação elementar de caracteres (formatos SGML, XML);

Na verdade, SGML e XML são, antes de tudo, linguagens que contêm um conjunto de regras, uma sintaxe, palavras-chave para criar documentos estruturados; quando um documento foi estruturado pela linguagem XML.

Formato digital: a representação de um objeto digital codificada como bytes, a qual define regras sintáticas e semânticas que permitem o mapeamento ou correspondência de um modelo de informação a uma sequência de bits e vice-versa. Na maioria dos contextos, o termo formato digital é usado indistintamente com conceitos relacionados a arquivos digitais, como formato de arquivo, envelope de arquivo, codificação de arquivo, etc. Também é conhecido como apresentação digital. (Barnard, A y Voutssas, J, 2014, p. 122).

Linguagem de marcação: um sistema de codificação legível por máquina, assim como suas regras associadas, que são utilizadas para descrever a estrutura lógica, distribuição, forma de exibição e estilo de um determinado documento digital. Existem várias linguagens de marcação com diferentes regras, propósitos e alcances: por exemplo, o HTML, o SGML e o XML. (Barnard, A y Voutssas, J, 2014, p. 142).

Linguagem de Marcação Padrão Generalizada (SGML): Linguagem de marcação padrão internacional ISO 8879:1986 utilizada para a definição formal de todo tipo de documentos, de modo que os torna independentes do dispositivo, sistema e programa com os quais foram criados. (Barnard, A y Voutssas, J, 2014, p. 143).

Linguagem de marcação extensível XML: linguagem de marcação padrão internacional desenvolvida pelo World Wide Web Consortium, ou W3C. XML é uma versão de SGML, projetado especialmente para os documentos da web. Permite que os designers criem suas próprias etiquetas, facilitando a definição, transmissão, validação e interpretação de dados entre programas e entre organizações. XML permite realizar funções que não podem ser feitas com HTML. (Barnard, A y Voutssas, J, 2014, p. 143).

Capítulo 2. Terminologia

A terminologia do capítulo é um pouco técnica. No entanto, é absolutamente necessário ser preciso sobre os diferentes termos que serão usados nesta seção.

Um formato, em seu sentido amplo, é uma forma padrão pela qual a informação é codificada para seu armazenamento em um arquivo de computador.

O formato permite definir:

- Quanto ao suporte físico, falar-se-á então do formato do suporte; nesse caso, serão especificadas as características físicas, por exemplo: A4 que é um formato de papel cujas dimensões são 21cm de largura x 29,7 cm de comprimento,
- As características lógicas de organização da informação, então falaremos do formato de dados, conceito de grande relevância para a presente seção,
- O conjunto de características físicas e lógicas que podem estar aninhadas, por exemplo (VHS, Discos Compactos, fitas magnéticas), situação pouco favorável para a preservação.

Nesta seção será apresentado uma série de exemplos concretos sobre os formatos. O formato terá múltiplas características, algumas delas são fundamentais para saber se o formato será adequado ou não do ponto de vista da preservação digital:

- **Formato fechado:** um formato fechado tem uma estrutura que não é de acesso público, exceto aqueles que a publicaram.
- **Formato normalizado:** um formato normalizado será considerado normalizado se estiver conforme uma norma emitida por uma organização ou entidade de normalização, como (ISO, AFNOR).

Atenção: um padrão que descreve um formato de dados pode ser apenas um contêiner dentro do qual devem estar inseridos os elementos que podem ou não ser normalizados, ou até mesmo ser privados.

- **Formato proprietário:** é um formato criado por uma empresa ou por um proprietário privado que detém os direitos de propriedade intelectual ou os direitos autorais correspondentes, por exemplo, PDF, TIFF, GIF...;

Aqui podem ser apresentados dois cenários:

I. O formato é proprietário, mas foi publicado; nesse caso, seu proprietário especificou os usos que estão autorizados, por exemplo, o XML do Microsoft Open Office, como DOCX e XLSX.

II. O formato é proprietário e não foi publicado (por exemplo, os arquivos produzidos pela Microsoft: DOC, XLS, PPT, formatos de imagem RAW e PSD do Adobe Photoshop).

- **Formato caseiro ou formato de projeto**: é um formato de dados definido especificamente por um aplicativo caseiro ou por um projeto dentro de uma empresa.
- **Formato público:** Trata-se de um formato cujas especificações foram publicadas e são acessíveis a todos sem restrição; isso não significa que o uso desse formato possa ser feito sem restrições.
- **Formato aberto:** Formato publicado e livre de direitos, impulsionado pela comunidade, sem restrições de uso e aplicação; é o caso de formatos definidos pelo consórcio W3C (por exemplo, HTML, PNG).

Formato padronizado: um formato será considerado padronizado se estiver em conformidade com um padrão.

Capítulo 3. A codificação

Tínhamos apresentado algumas generalidades sobre a codificação na seção 3. Aqui voltamos ao tema de forma mais detalhada, exemplificando as diferentes categorias de informação que podemos codificar:

- os caracteres, no mundo de hoje, são os caracteres de todas as línguas do mundo que devem poder ser codificados; é também nossa capacidade de produzir documentos multilíngues com técnicas de codificação compatíveis,
- os números de qualquer natureza e precisão. Algumas formas de codificação de números serão mais adequadas para a redução do espaço de armazenamento ocupado e para facilitar a manipulação nos cálculos,
- a codificação das cores seria outro exemplo.

ATENÇÃO

Não somos obrigados, para continuar com a seção, a entrar em detalhes, mas é conveniente sempre analisar o tamanho dos arquivos a serem usados.

3.1. Torre de Babel da codificação de caracteres

Até os anos 80, cada fabricante de computadores usava sua própria codificação.

EXEMPLO

Control Data Corporation utilizava o «Display Code», codificação de apenas 6 bits.

A IBM, por sua vez, criou o código EBCDIC (Extended Binary Coded Decimal Interchange Code), um modo de codificação de caracteres de 8 bits.

Esta multiplicidade de códigos causava múltiplas incompatibilidades entre os diferentes computadores.

COMPLEMENTO: CÓDIGO ASCII

O código ASCII (American Standard Code for Information Interchange) é um código de 7 bits que permite representar todos os caracteres anglo-saxões utilizados por uma série de fabricantes de computadores. Também é a variante estadunidense da norma de codificação de caracteres ISO/IEC 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

As representações da letra A em maiúsculas no Display code, ASCII e EBCDIC

Todos esses códigos sofreram variações ao longo do tempo, mas nenhum deles permitia representar caracteres latinos, gregos, cirílicos, etc.

Na década de 1990, ISO padronizou e ampliou o código ASCII, criando a norma ISO 8859:

A codificação é feita sistematicamente em 8 bits.

- Os primeiros 128 caracteres são os de ASCII,
- Os seguintes 128 são específicos do idioma.

16 versões dessa norma foram criadas para todas as línguas europeias, o hebraico, o cirílico, o árabe e algumas outras, tendo sido revisadas em 2020, permanecendo atualmente vigentes.

COMPLEMENTO: AS 16 VERSÕES DA NORMA DE CODIFICAÇÃO ISO 8859

ISO 8859-1 (latim-1 ou europeu ocidental)	É a parte mais utilizada de ISO 8859, abrange a maior parte das línguas europeias ocidentais: alemão, inglês, inglês basco, catalão, dinamarquês, escocês, espanhol, feroês e finlandês (parcialmente), francês (parcialmente), islandês, irlandês, italiano, neerlandês (parcialmente), norueguês, português, reto-românico e sueco, algumas línguas europeias do sudeste (albanês), assim como as línguas africanas (afrikaans e suaíli). O símbolo do euro e a letra maiúscula, que estavam faltando, estão na versão revisada ISO 8859-15 (latim-9). O conjunto de caracteres correspondente à ISO-8859-1 aprovado pela IANA, é a codificação que, por padrão, era usada nos documentos HTML ou documentos transmitidos por mensagens MIME, como as respostas HTTP quando o tipo de meio do documento é «text» (por exemplo, documentos «text/html»).
ISO 8859-2 (latim-2 ou europeu central)	Idiomas da Europa Central e Oriental baseados em um alfabeto romano (bósnio, croata, o polonês, tcheco, eslovaco, esloveno e húngaro).
ISO 8859-3 (latim-3 ou europeu do sul)	Para idiomas como turco (no início), maltês e esperanto; substituído pelo ISO 8859-9 para o turco e por Unicode para o esperanto.
ISO 8859-4 (latim-4 ou europeu do norte)	Estoniano, letão, lituano, groenlandês e o sami.
ISO 8859-5 (cirílico)	A maioria das línguas eslavas com alfabeto cirílico, incluindo o bielorrusso, o búlgaro, o macedônio, russo, sérvio e ucraniano (parcialmente).
ISO 8859-6 (árabe)	Abrange os caracteres mais comuns do árabe. Requer um mecanismo de renderização que suporte visualização bidirecional e análise contextual.
ISO 8859-7 (grego)	A língua grega moderna (ortografia monotônica).
ISO 8859-8 (hebraico)	O alfabeto hebraico moderno, tal como é usado em Israel.
ISO 8859-9 (latim-5 ou turco)	Similar à ISO 8859-1, onde as letras islandesas pouco são utilizadas, são substituídas por letras turcas. Também é utilizado para o curdo.

ISO 8859-10 (latim-6 ou nórdico)	Reordenamento do Latim-4. Considerado mais útil para as línguas nórdicas. As línguas bálticas utilizam com mais frequência o Latim-4.
ISO 8859-11 (tailandês)	Contém a maioria dos glifos necessários para a língua tailandesa.
ISO 8859-12	Originalmente abrangia os idiomas celtas. ISO 8859-12 foi configurado mais tarde para Latim/ Devanagari, mas foi abandonado em 1997, durante a 12ª reunião do ISO/IEC JTC 1/SC 2/ WG 3 na Grécia. A língua Celta é incluída na ISO 8859-14.
ISO 8859-13 (latim-7 ou báltico)	Adiciona alguns caracteres adicionais para as línguas bálticas que faltavam no Latim-4 e (latim-6).
ISO 8859-14 (latim-8 ou celta)	Abrange línguas celtas como o irlandês (ortografia tradicional), o gaélico escocês, o manês (língua extinta) e o bretão (algumas ortografias antigas).
ISO 8859-15 (latim-9)	Revisão do 8859-1 que abandona alguns símbolos pouco usados, substituindo-os pelo símbolo do euro € e pelas letras Š, š, Ž, é, Œ, y Ÿ, o que completa a cobertura do francês, do finlandês e do estoniano.
ISO 8859-16 (latim-10 ou europeu do sudeste)	Previsto para o albanês, croata, húngaro, italiano, polonês, romeno e esloveno, mas também finlandês, francês, alemão e irlandês (nova ortografia). Esta norma aposta mais nas letras que nos símbolos. O símbolo de moeda é substituído pelo símbolo do euro.

As normas de codificação devem evoluir, pois a linguagem evolui e surgem novas necessidades.

Exemplo:

O símbolo € e os Œ, œœet Ÿ que faltavam para a escrita do francês, foram adicionados à norma 8859-15, que constitui uma revisão do ISO Latim-1 (8859-1).

Como o número de posições em um byte está limitado a 256, um número de símbolos pouco usados foi abandonado em benefício de alguns novos.

COMPLEMENTO: DA ISO LATINA-1 À ISO LATINA-15: DIFERENÇAS ISO 8859-15 -- ISO 8859-1

Posição	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 sua vez, o Windows utiliza a codificação Windows-1252, idêntica à ISO 8859-1, exceto no intervalo 80-9F.

Conclusão

Mas então, como escrever um texto misto francês - grego?
Em geral, como codificar documentos multilíngues?
Como levar em conta as línguas asiáticas?
Como levar em conta o significado da escrita?

3.2. Codificação universal de caracteres

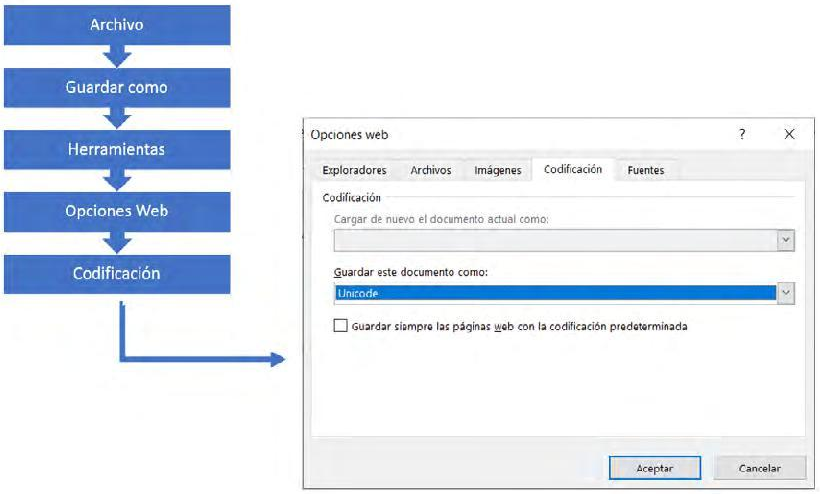
Em 1991 foi constituído o Consórcio Unicode para responder a essas perguntas e tentar solucionar definitivamente o problema da representação de diferentes caracteres de idiomas em um computador. Criando assim um padrão universal de caracteres (Unicode).

Durante os anos 90, o Consórcio Unicode e o Comitê Técnico ISO/IEC JTC 1, Tecnologias da Informação, por meio de seu subcomitê SC2 (Conjuntos de caracteres codificados), uniram seus esforços, já que nessa época existiam duas iniciativas para a codificação universal de caracteres: o padrão Unicode e a norma internacional ISO 10646.

Desde essa data, o Consórcio Unicode e a ISO desenvolveram o padrão Unicode e ISO/IEC 10646 em conjunto. O repertório, os nomes dos caracteres e os pontos de código Unicode Versión 2.0 coincidem exatamente com os da ISO/IEC 10646-1:1993, incluindo suas primeiras sete emendas publicadas. Depois que o Unicode 3.0 foi publicado em fevereiro de 2000, os caracteres novos e atualizados correspondentes foram incorporados à ISO/IEC 10646-1: 2000. Em 2003, as partes 1 e 2 da ISO/IEC 10646 foram combinadas em uma única parte, que desde então teve várias modificações que adicionam caracteres ao padrão em sincronização aproximada com o padrão Unicode, sendo sua versão mais recente a ISO/IEC 10646:2020 Information technology — Universal coded character set (UCS), pelo seu nome em inglês.

3.3. Onde configurar a codificação?

A forma de especificar a codificação que se deseja usar dependerá do software. No Microsoft Word 2019, é exibido da seguinte forma:



Como especificar a codificação no Word 2019.

Observa-se que, caso não seja configurada nenhuma especificação de codificação, o Word trabalhará com uma codificação padrão do Windows, que não é uma codificação padronizada.

Nos arquivos XML, dos quais falaremos mais adiante, encontra-se no início de cada arquivo a declaração que indica a codificação utilizada:

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

3.4. Codificação numérica

Para poder representar números, são atribuídos valores às posições dos bits.

COMPLEMENTO: CASO DE NÚMEROS INTEIROS - REPRESENTAÇÃO BINÁRIA

O número inteiro é decomposto em potências de 2 e atribui-se 0 ou 1 a

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

183 poderá estar representado por

10110111 ou 11101101 dependendo da direção que será usada para ler essa pequena sequência de bits.

No primeiro caso, os bits de peso mais alto (os atribuídos às potências maiores de 2) à esquerda,

No outro, colocamos os bits de peso mais alto à direita:

- O primeiro enfoque se chama Big Endian (Big Endian em inglês) e é utilizado

nos processadores Intel (por exemplo, no Windows).

- O segundo enfoque se chama pequeno-boutista (Little Endian em inglês) e é utilizado pelos processadores Motorola (por exemplo, no MacOS).

Dependendo do tamanho do número inteiro, são necessários mais ou menos bits para sua representação:

- com 8 bits, representam-se números inteiros positivos de 0 a 255.
- com 16 bits chegaremos a 65536
- com 32 bits até 4 294 967 296
- com 64 bits até 1 844 674 073 709 551 616

Na verdade, os valores são menores porque é necessário reservar um bit para o sinal.

No caso dos números inteiros - representação chamada codificada

Pode-se utilizar outro método muito simples para representar um número inteiro:
O número 183 pode ser representado codificando sucessivamente:

- o caractere «1» da codificação universal de caracteres (aqui idêntica à ISO Latin-1) (0110001)
- o caractere «8» da mesma codificação (0111000)
- o caractere «3» da mesma codificação (0110010)

Essa representação apresenta uma vantagem e uma desvantagem:

- la ventaja es que este valor será inmediatamente legible en cualquier editor de texto que interprete los caracteres;
- a desvantagem é que, para representar «183» de forma codificada, são necessários 3 bytes, ou seja, 24 bits, enquanto que apenas 8 bits são suficientes para representar esse número em forma binária; essa desvantagem será marginal em muitos casos, mas se tornará problemática quando se tratar de grandes volumes de dados; será mais rápido e mais econômico armazenar 1 GB em forma binária do que 3 GB em forma codificada.

Portanto, podemos dizer que o número inteiro positivo 183 pode ser representado por:

- a sequência de bits 10111 (modo binário),
- a sequência de bits 0110001 0111000 0110010 (modo codificado)

Essas sequências de bits só poderão ser interpretadas corretamente se dispusermos de uma informação que especifique o modo de representação do número, informação que, por sua vez, constituirá uma parte da Informação de representação, segundo o modelo OAIS.

Codificação dos números racionais e, de forma mais geral, dos números reais:

Em informática, também falaremos de números em ponto flutuante.

Em teoria, um número real está formado por 4 elementos:

- a mantissa (número inteiro positivo),
- o sinal do número real,
- o exponente,
- o sinal do exponente.

Mais uma vez, podemos definir uma representação chamada binária e uma representação chamada codificada. A representação codificada seguirá as mesmas regras que para os números inteiros.

O número -523,12 pode ser codificado com 7 caracteres: o sinal «-», o separador «vírgula» entre a parte inteira e a parte fracionária, e os números necessários para compor o número.

O número 1,6327-4 poderá ser codificado, na prática, na forma +1,6327E-04 (o caractere «E» serve para especificar que o número seguinte é um expoente). Serão necessários 11 bytes nesse caso.

A representação binária dos números reais será utilizada com mais frequência para realizar cálculos. Um artifício técnico de normalização entre a mantissa e o expoente permite que o expoente seja sistematicamente positivo e, portanto, não seja necessário armazenar o sinal do expoente.

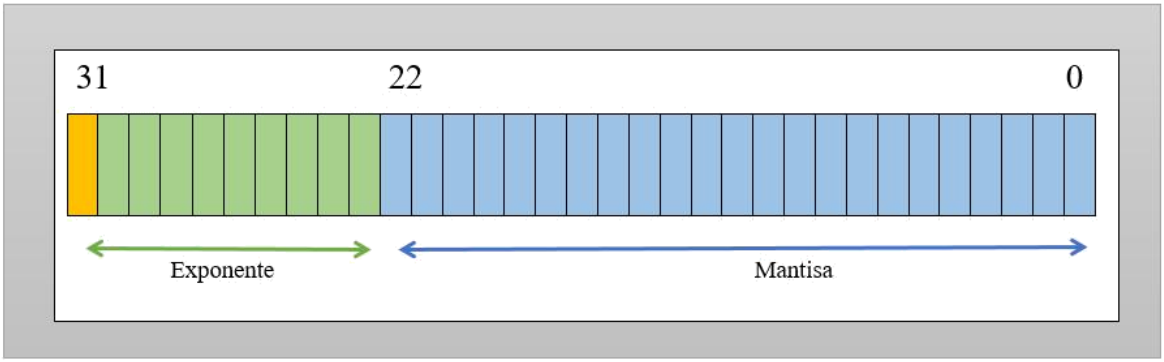
Em geral, serão utilizadas, dependendo dos computadores e das necessidades de precisão, representações com comprimento de 32, 64 ou, às vezes, 128 bits. Isso significa que o comprimento acumulado da mantissa, do expoente e do sinal deve corresponder a essas longitudes.

Mais uma vez, existem várias convenções diferentes dependendo dos sistemas

operacionais e dos fabricantes. Um destes convênios está padronizado e seu uso é

recomendado: Trata-se da norma 754-2008 IEEE Standard for Floating-Point Arithmetic.

Segundo esse padrão, um número de ponto flutuante de precisão simples é armazenado em 32 bits: 1 bit para o sinal, 8 bits para o expoente e 23 bits para a mantissa.



Representação IEEE de um número real de 32 bits. Elaboração própria a partir de PIAF.

Um número de ponto flutuante de dupla precisão é armazenado em 64 bits: 1 bit para o sinal, 11 bits para o expoente e 52 bits para a mantissa.

3.5. Outras codificações

Para a codificação de cores, trata-se também de definir um padrão que permita especificar a cor de um pixel em uma imagem, de um caractere em um texto, etc.

O pixel é a menor unidade direcional da tela.

Existem várias possibilidades:

- Codificação de uma paleta de 256 cores: nesse caso, um byte será suficiente para definir a cor,
- Codificação de 3 bytes (24 bits): 8 bits são dedicados à cor primária vermelha, 8 bits à cor primária verde e 8 bits à tonalidade primária azul,
- Codificação de 4 bytes (32 bits): 24 bits são utilizados como na codificação anterior, e o último byte é usado, por exemplo, para codificar informações sobre transparência.

Há códigos específicos para o som, outros para o vídeo (visto como um conjunto de imagens).

Capítulo 4. Quais são os critérios para avaliar os formatos?

Quais são os critérios que nos ajudarão a avaliar um formato de arquivo em relação à sua preservação a longo prazo?

Esses critérios deverão permitir:

- identificar os formatos que facilitarão a preservação da informação,
- eliminar os formatos que apresentem dificuldades graves a curto ou longo prazo.

Garantir o acesso aos dados pelo maior tempo possível implica poder conservar a informação em seus formatos originais e dispor dos meios para migrar para outro formato, se necessário. Mas cumprir com essas garantias não é suficiente; é preciso também conservar as propriedades significativas que permitam representar a informação.

Dada a rapidez da evolução técnica, é muito difícil prever a solução a ser adotada. Portanto, recomenda-se gerar os dados em formatos que sigam padrões, para que, no futuro, seja muito mais fácil buscar possíveis soluções.

A partir desse paradigma, podemos avaliar dois critérios principais: abertura e independência, aos quais podem ser adicionados critérios complementares.

4.1. Abertura: formatos aberto ou público

Um formato público ou publicado deve ter uma documentação completa e acessível. Essa documentação deve estar validada e ser suficientemente detalhada para permitir a escrita do programa, a leitura dos dados ou a conversão para outro formato, o que constitui uma garantia essencial de segurança para o futuro.

É preferível que o formato seja aberto, ou seja, livre de direitos, mas esse ponto não é determinante; no entanto,

recomenda-se verificar as restrições legais, pois o uso de determinados formatos pode ter certas condições, como ser pago.

Um formato proprietário amplamente difundido (como PDF ou TIFF) será preferível a um formato aberto pouco utilizado.

Quanto mais um formato é difundido, mais ferramentas são desenvolvidas para explorá-lo. A ampla difusão de um formato, no entanto, não é um critério que garanta por si só que esse formato seja adequado para a preservação a longo prazo.

EXEMPLO:

o formato DWG (Drawing, traduzido como desenho) é um formato fechado utilizado pelo software AutoCAD, propriedade da Autodesk.

Esse formato é muito utilizado por topógrafos, geógrafos e arquitetos. Mas apesar do seu uso mundial massivo, a preservação de dados em formato DWG apresenta graves dificuldades.

ATENÇÃO

Tenha cuidado com os chamados formatos “contêiner” ou “metarquivos”, que especificam uma estrutura, mas permitem, ao mesmo tempo, algoritmos de compressão, como no caso dos formatos de imagem, áudio e vídeo.

Na verdade, o formato pode ser livre de direitos, mas o algoritmo de compressão pode ser aberto e livre de direitos ou não, podendo até mesmo ser pago, dependendo da escolha feita.

4.2. Independência

Para ser adequado à preservação a longo prazo, um formato deve ser totalmente independente. Essa independência deve ser caracterizada por:

- em relação a outros formatos: alguns formatos podem parecer abertos, mas utilizam outros formatos que podem ser fechados ou sujeitos a patentes, o que limita o alcance de seus usos; também podem utilizar outros elementos, como conjuntos de caracteres, que, em alguns casos, podem ser proprietários;
- com relação aos sistemas operacionais: quando os formatos de dados estiverem associados a um sistema operacional específico, haverá uma restrição de acesso;

- plano econômico: mesmo no caso de formatos abertos, os custos de desenvolvimento das ferramentas devem ser razoáveis, para permitir que as organizações ou uma comunidade restrita de usuários assumam seu desenvolvimento;
- plano material: trata-se de garantir que o formato escolhido não esteja vinculado a um dispositivo ou meio de armazenamento específico que não seja controlado.

4.3. Outros critérios a considerar

Quando várias soluções satisfazem os critérios mencionados, podem ser levados em consideração outros critérios como:

- a disponibilidade e o custo das ferramentas e facilidades para criação de dados, assim como das ferramentas para transformação de formatos e apresentação de dados,
- a possibilidade de verificar automaticamente se um arquivo está em conformidade com as especificações do formato e também respeita as normas de uso e as restrições definidas para a preservação,
- evitar o uso de formatos desnecessariamente volumosos,
- complexidade: um formato simples é preferível a um formato completo
- a estrutura do formato: quanto mais o conteúdo e o estilo se misturarem no formato, mais difícil será recodificar um sem modificar o outro ou adaptar outro estilo ao mesmo conteúdo,
- a disponibilidade e o potencial de desenvolvimento de serviços de valor agregado, como a extração de metadados ou a conversão para formatos de divulgação

Naturalmente, é difícil cumprir todos esses critérios complementares. Podem até mesmo ser contraditórios entre si: um formato pode ser mais simples, mas resultar em volumes de dados maiores. Portanto, devem ser analisados esses requisitos em função das características e do contexto do Arquivo.

ATENÇÃO

Quanto menor for o número de formatos de documentos aceitos e gerenciados pelo Arquivo, menor será o risco de preservação da informação.

4.3. Recomendações existentes

Atualmente, existem múltiplas propostas a nível mundial que oferecem recomendações sobre quais formatos deveriam ser utilizados ou as características que estes deveriam cumprir para garantir a preservação a longo prazo. A seguir, alguns exemplos:

- **Norma NTC-ISO 19005-1:2020 Gestão de documentos. Formato de arquivo de documento eletrônico para preservação a longo prazo. Parte 1: Uso do PDF 1.4 (PDF/A-1):** Esta parte da norma NTC-ISO 19005, especifica

como se utiliza o Formato de Documento Portátil (PDF) 1.4 para a preservação a longo prazo de documentos eletrônicos, o que é aplicável a documentos que contêm combinações de caracteres, dados raster e vetoriais.

- **Norma UNE-ISO 15489-1:2016. Informação e documentação. Gestão de documentos. Parte 1 e 2:** O qual estabelece um marco de preservação sistêmica para documentos por meio da definição de considerações de ordem geral (Parte 1), assim como uma metodologia de implementação (Parte 2) para um sistema de gestão de documentos.
- **Projeto Interpares 3:** O qual busca, entre outras coisas, criar modelos de avaliação capazes de medir o sucesso das soluções de preservação que se proponham e implementem, assim como emitir diretrizes voltadas aos requisitos de preservação aplicáveis a tipos específicos de documentos de arquivo.
- **UNE-ISO 23081-1:2018. Informação e documentação. Processos de gestão de documentos. Metadados para a gestão de documentos. Parte 1: E**stabelece um marco para a criação, gestão e uso de metadados para a gestão de documentos, e explica os princípios pelos quais devem se reger.
- **UNE-ISO 23081-2:2021 Informação e documentação. Processos de gestão de documentos. Metadados para a gestão de documentos. Parte 2: Elementos conceituais y de implementação.:** Esta especificação técnica proporciona um marco para definir elementos de metadados consistentes com os princípios e as considerações de implementação descrita na Norma supra citada. Apresenta uma abordagem lógica aos metadados para a gestão de documentos nas organizações, modelos conceituais de metadados e uma visão geral do conjunto dos diferentes tipos de elementos de metadados válidos para qualquer ambiente de gestão de documentos.
- **UNE-ISO/TR 15801:2019 IN Gestão de documentos. Informação armazenada eletronicamente. Recomendações sobre confiabilidade e fiabilidade:** Define os princípios que devem reger a gestão da informação, considerando os seguintes aspectos: políticas de gestão da informação, dever de custódia, procedimentos e processos, tecnologias capacitadoras e trilha de auditoria.
- **Princípios e requisitos funcionais para documentos em ambientes de escritório eletrônico:** Publicado pelo Conselho Internacional de Arquivos (ICA), seu propósito é definir globalmente alguns princípios e requisitos funcionais harmonizados para o software utilizado para criar e gerenciar documentos eletrônicos em ambientes de escritório.
- **MoReq. Modelo de requisitos para a gestão de documentos eletrônicos de arquivo:** Descreve um modelo de requisitos para a gestão de documentos eletrônicos de arquivo; foca nos requisitos funcionais dos sistemas de gestão de documentos eletrônicos de arquivo.

A lista anterior não pretende ser exaustiva, ao contrário, é um convite para analisar mais detalhadamente a multiplicidade de propostas que existem a nível mundial.

COMPLEMENTO

É recomendável monitorar e identificar os formatos de arquivos adequados para o arquivo digital, para o qual pode criar uma lista de formatos de arquivo.

Em primeiro lugar, é preciso identificar que tipo de conteúdo digital (existente ou futuro) precisa ser preservado, pois isso vai definir a escolha dos formatos. No entanto, não é necessário começar do zero, podendo entrar em contato com organizações semelhantes.

Existem organizações e instituições que publicaram sua lista de formatos de arquivo, por exemplo, a Biblioteca do Congresso dos Estados Unidos publicou sua lista atualizada em 2023: https://loc.gov/preservation/ resources/rfs/RFS%202022-2023.pdf (material disponível em inglês).

Durante a fase de elaboração deste insumo, lembre-se de ser realista quanto às capacidades de preservação do arquivo digital e se elas podem ser cumpridas conforme necessário; talvez sejam identificados entre os formatos aceitáveis e os preferidos.

Além disso, devem ser consideradas revisões posteriores que reflitam a tecnologia atual e as boas práticas, de acordo com a pesquisa e análise realizadas.

4.4. Registros de formatos

A coleta de informações e documentação sobre os formatos digitais e os programas de computador que permitem criar ou ler dados organizados segundo esses formatos é um trabalho complexo (características, tipo, disponibilidade da documentação, direitos de propriedade aplicáveis), diante da constante mudança e evolução a que estão sujeitos os diferentes formatos, exigindo recursos humanos e econômicos significativos.

Portanto, é ilusório imaginar que uma instituição possa acompanhar por si mesma a evolução dos formatos digitais que gerencia.

O objetivo dos registros de formatos é trabalhar de forma conjunta na coleta, atualização e identificação de informações por parte de uma comunidade de usuários.

EXEMPLO:

A principal iniciativa e que está atualmente em funcionamento:

- PRONOM no Reino Unido (https://www.nationalarchives.gov.uk/PRONOM/)

O qual possui um repositório de pesquisa no GitHub que permite aos usuários colaborar com o projeto PRONOM e contém guias úteis sobre como os usuários podem se envolver na pesquisa de formatos, estando próxima de sua atualização para a versão 109.

As informações coletadas para cada formato em um registro incluem, entre outras coisas:

- os nomes pelos quais um formato é conhecido e suas variantes; por exemplo, PDF, Adobe PDF, Portable Document Format
- Sua extensão: por exemplo, .pdf
- as especificações do formato
- os autores, titulares de direitos, encarregados da manutenção
- relações com outros formatos derivados, versões
- os sistemas, serviços e ferramentas para a criação, leitura e validação de documentos conformes a esse formato

Outra iniciativa útil nessa área é o site informativo da Biblioteca do Congresso dos Estados Unidos, que oferece um conjunto de informações, publicações e recursos sobre formatos e sua possível preservação. O site apresenta uma descrição dos formatos classificados por tipo (texto, imagem, áudio, vídeo).

Capítulo 5. Os formatos de dados

Até agora foi falado sobre como representar entidades simples como números ou caracteres.

Agora trata-se de representar objetos digitais mais complexos que podem conter texto, imagens, gráficos, som e vídeo. Essas diferentes categorias de informação também podem ser coletadas, organizadas e combinadas em documentos chamados multimídia.

A representação desses objetos apresenta outros problemas que se sobrepõem aos anteriores.

5.1. O papel especial de XML

Origem

Desde a década de 1970 e início da década de 1980, a IBM analisava a possibilidade de armazenar textos em computador. Três de seus engenheiros desenvolveram uma linguagem de marcação para separar as instruções de estilo do conteúdo dos documentos; Charles Goldfarb, Edward Mosher e Raymond Lorie deram a essa linguagem o nome baseado em suas iniciais: GML.

Os trabalhos da IBM são retomados pela Organização ISO e levam à criação da norma ISO 8879/1986, batizada de SGML, que significa Standard Generalized Markup Language, em inglês.

Conceitos

O SGML introduz os conceitos essenciais: conteúdo e formato separando ambos:

- Um mesmo conteúdo pode ser utilizado com diferentes estilos

  - papel

  - tela

  - telefone celular

- o formato do documento: Faz referência à marcação, o que permite atribuir uma verdadeira semântica aos diferentes elementos da estrutura (título do documento, título do capítulo, glossário, nota de rodapé, etc.).

De SGML a XML

SGML como linguagem de marcação e descrição de documentos, era tão complexa que só podia ser utilizada por especialistas (técnicos).

Em 1990, com a chegada da Web e da linguagem HTML (Hypertext Markup Language, uma linguagem construída com base nos princípios do SGML), ficou evidente que o SGML era muito complexo. Por isso, o World Wide Web Consortium (W3C) criou um grupo de trabalho para analisar o SGML e transformá-lo em uma versão com muito mais facilidade de implementação. Isso leva à criação do **Extensible Markup Language** (XML) 1.0 do 10 de fevereiro de 1998, por parte do W3C.

Princípios

XML significa em inglês Extensible Markup Language e é uma linguagem de descrição de documentos que não inclui nenhuma informação relativa ao design desses documentos. **XML é uma metalinguagem** que permite definir outras linguagens de marcas com objetivos diferentes. Por esse motivo, é chamado de "extensível". Portanto, XML não é realmente uma linguagem em particular, mas sim uma maneira de definir linguagens específicas que terão as seguintes propriedades:

- todos eles serão conformes ao XML, ou seja, compartilharão a mesma sintaxe, as mesmas regras e poderão ser manipulados por ferramentas genéricas;
- serão interoperáveis, ou seja, será possível criar documentos compostos nos quais vários idiomas XML poderão cooperar;
- serão etiquetados de maneira que evidenciem a estrutura do documento; o conteúdo e o estilo serão separados.

Portanto, qualquer documento XML pode ser:

- validado por um analisador sintático (também chamado de “parser”),
- Formatado: o qual pode ser realizado pelas linguagens CSS (Cascading Style Sheets: folhas de estilo em cascata) ou XSL-FO (extensible Stylesheet Language-Formatting Objects). CSS e XSLFO são dois padrões do W3C,
- Transformado sua linguagem de estilo, por exemplo, com XSLT (extensible Stylesheet Language Transformations).

Linguagens e formatos

Quando um documento foi estruturado em linguagem XML, conhece-se previamente o conjunto de regras de organização da informação dentro desse documento. Portanto, o XML pode ser considerado como um formato aberto e legível tanto por computadores quanto por humanos.

XML utiliza o set de caracteres de Unicode e permite o uso de diferentes codificações incluindo UTF-8 que é a codificação pré-determinada.

XML está padronizado por W3C.

Até hoje, existem centenas, talvez milhares, de linguagens XML dedicadas a aplicações empresariais específicas. Um grande número de formatos de dados e metadados nos mais diversos campos baseiam-se na sintaxe XML.

Alguns são bastante genéricos para ser amplamente utilizados:

- SMIL (Synchronized Multimedia Interfaz Language) para documentos multimídia,
- SVG (gráficos vetoriais escaláveis) para gráficos vetoriais em 2D,
- Xforms para formulários.

XML não é um padrão ISO, mas muitas das suas normas baseiam-se nele, muitas plataformas se beneficiam disso.

ATENÇÃO

É preciso acrescentar que, se quiser preservar documentos XML, é necessário validar esses documentos no momento do ingresso no repositório e, portanto, dispor dos modelos (templates) aos quais esses arquivos fazem referência, além de preservar também esses modelos. Pode ser uma DTD (Definição de tipo de documento), esquemas XML ou qualquer outro tipo de gramática padronizada.

A adoção de modelos como DocBook, TEI (Text Encoding Initiative), textML, ALTO, é essencial na perspectiva do arquivo.

XML, um ativo para a preservação dos documentos digitais

- XML é um formato aberto padronizado pelo W3 Consortium e utilizado mundialmente em diversos setores,
- XML é legível tanto para seres humanos quanto para computadores e, por isso, pode ser convertido sem dificuldade. É sintaticamente verificável,
- XML dissocia o conteúdo do estilo e permite assim associar diferentes estilos a um mesmo conteúdo,
- XML é interoperável e não depende da plataforma informática utilizada,
- XML depende apenas da codificação de caracteres, que é feita por UTF-8

5.2. Formatos de texto

Documentos de texto editável

Também conhecido como texto sem formato, texto simples (Plain Text), esses são documentos que não possuem nenhum tipo de formatação tipográfica e que podem ser lidos por qualquer ser humano.

É um formato aberto, cujo conteúdo dependerá da codificação de caracteres, existem muitas ferramentas para editar, manipular e converter esses documentos em múltiplos sistemas operacionais.

EXEMPLO:

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

Os documentos em texto plano não apresentam dificuldades quanto à sua preservação, pois a codificação utilizada é conhecida. No entanto, são documentos “pobres” já que contém texto sem estrutura e estilo.

Há uma série de formatos utilizados pelas suítes de escritório, os dois principais são:

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

Formato	Descrição
ODF - (Open Document Format)	O Formato ODF (OpenDocumentFormato) é um formato aberto baseado em uma linguagem normalizada de definição de esquema do documento RELAX NG, e inclusive construído sobre a linguagem XML. ODF foi padronizado pelo consórcio OASIS (Organization For The Avance da informação estruturada Standards) em 2005 e tornou-se a norma ISO 26300 em 2006 sendo sua última versão de 2015.

OOXML - Office Open XML	Os formatos oferecidos pela Microsoft Office como (.doc, .xls, .ppt (para as versões mais antigas) y .docx, .xlsx e .pptx (para as versões recentes 2007 em diante) foram formatos fechados, e só foram publicados no início dos años 2000. Esses formatos, assim como o pacote de software da Microsoft Office evoluíram lançando uma nova versão a cada dois anos desde 1990. **Atenção** **Muito importante!** **A retrocompatibilidade, ou seja, a capacidade de ler um arquivo criado com uma versão anterior por parte de uma versão nova na suite Office, não está garantida após 10 anos.** A Microsoft desenvolveu seu próprio formato aberto de arquivo, o MS-OOXML o qual está baseado em XML para documentos de escritório. Abrange os documentos do processador de texto, planilhas, apresentações, gráficos, diagramas, figuras. MS-OOXML foi adotado pela primeira vez como padrão em 2006 por parte da ECMA, uma organização privada de padrões internacionais supostamente como um padrão aberto. Em 2008, a Organização Internacional de Normalização (ISO) também aprovou o MS-OOXML, como Padrão Aberto internacional sob a Norma ISO/IEC 29500.

A situação em relação aos formatos de documentos de escritório mudou drasticamente nos últimos anos.

ODF parece ser uma das opções para o arquivamento de longo prazo de documentos editáveis, mas não podemos prever o que acontecerá no futuro.

A realidade mostra que se deve ter cautela nesse aspecto.

Documentos de texto não editável

O formato PDF (Portable Document Format) e sua versão PDF/A dedicada ao arquivo

Formato	Descrição
PDF 1.7 e 2.0	PDF é um formato proprietário publicado. Pertence à empresa Adobe. É um formato contêiner. Este permite conter outros tipos de formato de dados, como imagens, som, vídeo, etc. Há muitas ferramentas para manipular esse formato. A política do Adobe é distribuir grátis as ferramentas de leitura e vender as ferramentas de criação. O fato de ser um formato contêiner obriga a verificar rigorosamente que os arquivos PDF destinados à preservação estejam construídos unicamente com os elementos previstos. O formato pode incluir metadados em formato XMP (Extensible Metadata Platform). Em julho de 2008, a versão 1.7 de PDF tornou-se a ISO 32000-1 2008 (Gestão de documentos - Formato de documento portátil - Parte 1: PDF 1.7), sendo sua última revisão em 2018. Por sua vez, em 2020, a versão 2.0 tornou-se a ISO 32000-2:2020 (Gestão de Documentos - Formato de Documento Portátil - Parte 2: PDF 2.0). Apesar de serem ambos formatos normalizados, isso não elimina a necessidade de tomar precauções indispensáveis já explicadas.
PDF/A	A versão 1.4 do PDF foi a base sobre a qual se definiu em 2005, a norma ISO 19005-1: Gestão de documentos. Formato de arquivo de documento eletrônico para a conservação a longo prazo. Parte 1: Uso do PDF 1.4 (PDF/A-1). Revisada em 2020, PDF/A contém uma série de restrições em comparação com o PDF mas integra todos os elementos necessários para o acesso do documento e, em particular, as propriedades de visualização. No ano de 2011 foi publicada a especificação PDF /A -2 (Formato de arquivo de documento eletrônico para a conservação a longo prazo. Parte 2), e está construída por volta da versão 1.7 do formato PDF e, por fim, no final do ano de 2012 foi publicada a terceira especificação o PDF /A-3, que permite a incorporação de arquivos em outros formatos. Normalmente ao utilizar esses padrões o resultado é um aumento do volume de um arquivo, mas em contrapartida, uma independência desses arquivos em relação às plataformas nas quais são utilizados.

ATENÇÃO

O formato PDF está destinado aos documentos não editáveis. Apresenta a vantagem de poder representar a apresentação original do documento de maneira fiel, garantia que nem sempre pode ser totalmente garantida pelos formatos ofimáticos editáveis. No entanto, não se deve acreditar que um documento PDF não possa ser modificado de forma mal-intencionada.

5.3. Formatos de imagem e gráficos vectoriais

Há dois tipos de formatos

- As imagens de pixel. Uma imagem em mapa de bits, imagem raster ou imagem de pixels, é uma estrutura ou arquivo de dados que representa uma grade retangular de pixels ou pontos de cor, denominada matriz, que pode ser visualizada em um monitor, papel ou outro dispositivo de exibição,
- As imagens vetoriais. Uma imagem vetorial ou gráfica é um tipo de imagem definida em um plano, conectada por linhas e curvas, dando como resultado formas baseadas em equações matemáticas. E devido a isso, se você aproxima ou afasta o zoom, as linhas, curvas ou pontos sempre permanecem suaves.

Formatos de imagem com descrição de pixels

Abreviação	Nome e estado	Características
GIF	• Formato de gráficos intercambiável • Formato proprietário publicado • Limitações relacionadas com sua patente	Possui uma patente de propriedade, o que poderia significar algumas restrições no seu uso.
TIFF	• Formato de arquivo de imagem etiquetada • Formato publicado • Propriedade da empresa Adobe • Sem licença de uso	É um formato contêiner: Define uma estrutura. Permite incluir os perfis ICC (International Color Consortium) no arquivo. O perfil ICC permite a gestão da cor independente das plataformas e dispositivos. A imagem poderá ser registrada segundo diferentes algoritmos de compressão, conforme a escolha. do usuário. As imagens também podem ser guardadas sem compressão.

		O sucesso desse formato deve-se a duas razões principais: • Permite registrar as imagens em preto e branco com o algoritmo ITU T6 que havia sido criado para a transmissão por meio do fax: esse algoritmo é muito eficaz e oferece taxas de relação de compressão elevadas sem perdas. • Esse formato também oferece a possibilidade de armazenar metadados técnicos muito complexos. TIFF é utilizado como formato de digitalização.
JPEG e JPEG2000	• JPEG (Joint Photographic Expert Group) é um formato publicado e aberto. • JPEG publicado como Norma ISO/IEC IS 10918-1, sua mais recente versão ISO/IEC 10918-1:1994/Cor.1:2005 • JPEG2000 é um formato publicado e aberto publicado como Norma ISO/CEI 15444-1, em seu mais recente revisão do 2019.	A norma JPEG descreve o algoritmo de compressão e a informação mínima para usá-lo. A razão do seu sucesso é que foi implementado de forma massiva como código aberto, além de ser admitido por todos os navegadores de Internet. Seu principal inconveniente é que utiliza um algoritmo de compressão com perda (ainda que essa taxa de compressão possa ser muito leve). JPEG não é propriamente um formato de arquivo, JPEG baseia-se no formato JFIF (formato de intercambio de arquivos JPEG), o qual define especificações complementares para o formato contêiner que contém os dados de imagem codificados com o algoritmo JPEG Com JPEG2000, a compressão é uma das melhorias importantes desse formato, em comparação com o JPEG, embora a qualidade seja semelhante
JBIG	trabalhad • Formato o por o Joint Bi-level Image especial istas Group, o qual desenvolveu a Norma ISO/ IEC IS 11544, a cual está vigente e conta com uma última revisão em 2021. • Formato publicado aberto. • Limitação de as patentes no algoritmo de compressão	JBIG utiliza um algoritmo de compressão sem perda. Pode também ser utilizado para a codificação a níveis de cinza e imagens coloridas com um número limitado de bits por pixel. O inconveniente é seu algoritmo de compressão, o qual está sujeito a uma patente de propriedade da IBM, Mitsubishi e Lucent. é provavelmente uma das razões da sua escassa difusão.

PNG Portátil network Graphics	• Formato publicado aberto • Padrão do W3C • Norma ISO/IEC 15948:2004 revisado e validado em 2021.	Formato que suporta imagens em escala de cinzas ou coloridas. É aceito pela maioria dos navegadores modernos.

Formatos com descrição vetorial

Abreviação	Nome e status	Principais características
SVG	• SVG Vector Scalable Graphics • Padrão aberto do W3C	Família de formatos baseados na linguagem XML abertamente documentados e utilizados para gerar gráficos vetoriais bidimensionais para seu uso na web, são desenvolvidos pelo World Wide Web Consortium (W3C).
DWG	• Abreviação de DraWinG (Desenho) • Formato fechado, propriedade da sociedade Autodesk, distribuidor do software Autocad	DWG é o formato de arquivo nativo patenteado para AutoCAD, uma das ferramentas de design por computador mais populares. DWG é um formato binário compacto que armazena e descreve o conteúdo de metadados e dados de design 2D e 3D.

5.4. Formatos de áudio e vídeo

Os formatos audiovisuais e sonoros constituem um âmbito tecnicamente complexo. No entanto, os critérios de avaliação dos formatos em relação à preservação digital são aplicados na íntegra.

Será feita uma distinção sistemática entre a especificação do formato, que define o modo de encapsulamento do conteúdo — ou seja, a organização da informação dentro de um contêiner — e a codificação propriamente dita da informação, que frequentemente utilizará um algoritmo de compressão de dados para reduzir seu volume.

Daremos algumas indicações gerais sobre uma seleção de formatos existentes.

ATENÇÃO

É comum separar os formatos de preservação, ou seja, aqueles que permitem conservar toda a informação e dos formatos de difusão, que são os que fornecem o conteúdo em resposta a uma solicitação por parte do consumidor.

Formatos de áudio

Os formatos de áudio são formatos "envolventes", chamados também formatos "contêiner". Dentro desses formatos, a informação de áudio pode ser codificada e comprimida de várias maneiras, isso com o objetivo de definir claramente como representar o áudio, por isso, é recomendável sempre especificar o tipo de codificação que será utilizado.

Há recomendações de formatos sonoros emitidas pela IASA (International Association of Sound and Audiovisual Archives). Lamentavelmente, não há equivalente para o vídeo.

Apresenta-se uma breve síntese dos principais formatos.

Abreviação	Nome e status	Principais características
Wave	• Formato publicado, propriedade da Microsoft.	Formato de arquivo para áudio, do tipo contêiner que pode incorporar um fluxo de bits de áudio com outros fragmentos de dados. Uma codificação de fluxo de bits comum é LPCM (modulação de código de pulso linear). Os projetos de preservação que utilizam o reformatamento geralmente utilizam uma das variantes de Broadcast WAVE. Completamente documentado. Formato patenteado desenvolvido pela Microsoft e IBM como parte do formato de arquivo de troca de recursos (RIFF) para Windows 3.1, com documentação disponível gratuitamente.
AIFF	● Audio Interchange File Format ● Formato publicado, propriedade da Apple	Formato de arquivo para som que envolve vários fluxos de bits de som, desde forma de onda sem comprimir até MIDI.
MP3	● MPEG Audio Layer 3 ● Algoritmo de compressão (ISO/CEI IS 11172-3 e ISO/ CEI IS 13818-3).	Formato de áudio dotado de um algoritmo de compressão capaz de reduzir grande quantidade de dados necessária para restituir áudio com perda de qualidade sonora aceitável para o ouvido humano. MP3 é o nome, para a codificação de áudio MPEG Layer III define-se em duas famílias de especificações ISO/ IEC (MPEG-1: 11172-3 e MPEG-2: 13818-3), sendo um padrão aberto. Não possui restrições para a estrutura do formato

OGG	● Formato publicado e aberto Formato aberto promovido pela fundação Xiph.org.	Esse formato deve ser utilizado com a codificação Vorbis, também definida por essa fundação. Vorbis é um algoritmo de compressão e descompressão de áudio digital, aberto e livre, de melhor rendimento em termos de qualidade e taxa de compressão que o formato MP3

Formatos de vídeo

As características de um formato de vídeo são complexas, por isso a escolha de um formato para preservação exige uma análise específica.

São propostos alguns elementos como base para uma reflexão introdutória e não exaustiva.

Abreviação	Nome e status	Principais características
MPEG	• Moving Pictures Expert Group • Formato aberto • Conjunto de normas ISO	Esse formato corresponde a uma família de normas ISO: • MPEG-1: os primeiros filmes na Internet (ISO/CEI 11172-1 a 5) • MPEG-2: a televisão digital atual • MPEG - 4: A televisão Digital Terrestre • MPEG-7, MPEG-21: futuras normas de composição de cenas, muito ricas em metadados. A mais utilizada atualmente é MPEG-4 que requer ser utilizada com uma codificação llamada H264. H.264, ou MPEG-4 AVC (Advanced Video Coding), é uma norma de codificação de vídeo desenvolvida conjuntamente pela UIT (Unión Internacional de Telecomunicaciones) e a ISO. A norma UIT-T H.264 e a norma MPEG-4, Parte 10 (ISO/ CEI 14496-10) são tecnicamente idênticas.
MJPEG 2000	• Motion JPEG 2000 • Formato publicado e aberto	Padrão aberto. Desenvolvido conjuntamente pelo Grupo de especialistas em imagens em movimento (MPEG), um grupo de trabalho dentro da estrutura do comitê ISO/IEC conhecido formalmente como ISO/IEC JTC 1/SC 29. Os quadros MJP2 são representados como entidades separadas codificadas com J2K_C (com perda ou sem perda).

OGG	• Formato publicado e aberto	Formato aberto promovido pela fundação Xiph.org. Esse formato deve ser utilizado com a codificação Theora, Esse é o modo de compressão de vídeo livre e sem patentes promovidas pela mesma fundação.
Matroska	• (Матрёшка ou Boneca russa) • Formato aberto	Matroska é um formato que pode agrupar em um mesmo arquivo várias faixas de vídeo, áudio, legendas e capítulos. Como em MPEG 4, é recomendado a codificação H264.
AVI	• Audio Video Interleave • Formato proprietário (Microsoft) • Formato contêiner publicado	Esse formato contêiner permite qualquer codificação. No modo não comprimido, os arquivos são muito pesados.

5.5. Formatos de arquivo produzidos por aplicações “caseiras”

Até agora, foram mencionados principalmente formatos abertos ou proprietários, que estão disponíveis no mercado e para os quais existem programas de escrita, leitura, conversão, etc.

Muitas empresas ou instituições desenvolvem as suas próprias aplicações de software para satisfazer suas necessidades. Os dados digitais produzidos por essas aplicações têm, portanto, um formato próprio.

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. Essa descrição deverá imperativamente:

- ser completa,
- ser precisa,
- ter sido validada rigorosamente.

ATENÇÃO

A validade e a completude desta descrição do formato de aplicação “caseira” são elementos determinantes para a preservação dos dados obtidos desta aplicação. Qualquer descumprimento desta documentação acarreta imediatamente um risco importante de perda ou interpretação errônea da informação preservada.

O detalhamento da forma como essa documentação deve ser apresentada está fora do escopo deste curso, mas é conveniente saber:

- Existem métodos de descrição formal. Estes métodos permitem apresentar uma descrição do formato que será interpretada tanto por pessoas quanto por programas de computador,
- As ferramentas permitem garantir a coerência entre um documento digital e a descrição formal do seu formato.

Capítulo 6. Este arquivo está no formato correto?

A extensão do nome do arquivo (.pdf, .doc, .xml...) permite identificar rapidamente o tipo de dados; no entanto, a extensão (especialmente em ambiente Windows) é insuficiente, pois vários tipos de dados compartilham a mesma extensão.

Por exemplo, para PDF podemos ter:

- 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 é um número que, situado no início de um arquivo, indica seu tipo aos programas que o processam. Esta técnica é utilizada por diferentes sistemas operacionais. como MacOS e Unix. Magic number oferece maior confiabilidade que a extensão oferece, cujo valor pode ser alterado voluntária ou facilmente.

Por exemplo, o número mágico para os arquivos PDF terá como valor %PDF-1.

Para os arquivos TIFF terá para magic number MM. * ou II* segundo como esteja constituído o arquivo.

No entanto, somente a análise completa dos formatos nos permitirá assegurar que o arquivo está conforme as especificações do formato a que se refere.

ATENÇÃO

A validação dos formatos dos arquivos que entram em um repositório constitui uma operação crítica. Se os arquivos transferidos pelo produtor não cumprirem plenamente as especificações do formato e, além disso, se as não conformidades não forem identificadas explicitamente nas ferramentas de leitura atuais, gera-se um risco importante para o repositório, que assume a responsabilidade de preservar a informação uma vez que ela é transferida e aceita.

Por exemplo, é comum que os arquivos PDF não cumpram a especificação publicada pela Adobe, sem que as anomalias sejam indicadas pela ferramenta de leitura gratuita Acrobat Reader. A validação dos formatos de entrada implica o uso e, quando necessário, o desenvolvimento de ferramentas de controle desses formatos e a aplicação de procedimentos muito rigorosos.

EXEMPLO:

A seguir, apresentam-se alguns exemplos que evidenciam as dificuldades causadas pela não validação dos documentos por parte dos repositórios documentais.

- Os arquivos HTML podem conter erros de sintaxe que deveriam ser rejeitados pela ferramenta de validação (repositórios); pelo contrário, se forem aceitos, alguns desses erros podem dificultar a migração do formato no futuro.
- A validação de arquivos XML implica referir-se a esquemas ou DTDs externas. Isso exige, portanto, ter previamente recuperado esses esquemas ou modelos e mantê-los no sistema de preservação,
- la validación de archivos de vídeo en formato MPEG es compleja debido a la gran permisividad de estructura de este formato,
- Até o momento, não existe uma solução para a conversão de arquivos gráficos (tipo AutoCAD).

Foram desenvolvidos diversos programas de validação de formatos.

EXEMPLO:

- JHOVE (JSTOR/Harvard Object Validation Environment) que permite validar a conformidade dos arquivos em relação com vários formatos, como AIFF (Audio Interchange File Format, formato de áudio de Apple), GIF, HTML, PDF, TIFF, JPEG (Joint Photographic Experts Group), XML, WAVE (formato de áudio de Microsoft), etc. JHOVE é um software livre sob licença GPL de GNU,
- DROID (Digital Record Object Identification) é uma ferramenta de código aberto proporcionada pelo Arquivo Nacional do Reino Unido. Baseia-se no registro de formatos PRONOM e vincula a identificação do formato aos documentos técnicos correspondentes disponíveis no registro.

Outra iniciativa útil é a do CINES, que disponibilizou online um serviço de validação de formatos utilizando como base o JHOVE, o DROID e outras ferramentas. Este serviço, chamado “Facile”, permite não precisar instalar nenhum software de validação de formato.

ATENÇÃO

A validação da informação digital é a chave da preservação a longo prazo, já que garante o acesso: dados e metadados.

Mas também pode se tornar fonte de múltiplas dificuldades técnicas, daí a necessidade do trabalho multidisciplinar, do compartilhamento de experiências e ferramentas.

Bibliografia

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 Gestão 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