Formatos de Arquivo para Preservação de Documentos Digitais[1]

Ernesto C. Bodê (PGCInf/UnB)
Miriam P. Manini (PGCInf/UnB)

Resumo. Esse artigo apresenta os resultados já obtidos numa pesquisa sobre o uso de formatos de arquivo adequados para a preservação digital por longos períodos. Utilizamos como fonte bibliográfica vários textos produzidos em centros internacionais de pesquisa sobre a preservação digital. Além da introdução do projeto de pesquisa, apresentamos os primeiros resultados no que se refere à concretização de temas importantes, como o próprio conceito de formato de arquivo. Apresentamos também resultados obtidos com relação às características desejáveis para uso de formatos de arquivo na preservação por longos períodos.

Palavras-chave: Documentos digitais. Preservação. Formatos de arquivo.

Abstract. The article presents some results from a research about the use of file formats for long-term digital preservation. The bibliographic source was made from many articles and papers at international research centers working with digital preservation. After an introduction of the research project's, we present the first results already built. We present relevant theoretical concepts, like the file format concept. Then we present some results related to expected and desired characteristics for the use of file formats for long-term preservation.

Keywords: Digital documents. Long-term preservation. File formats.

[1] Comunicação oral apresentada ao GT-08 - Informação e Tecnologia.

1 – Introdução e Justificativa

Entre tantas novidades boas e não tão boas, a contemporaneidade trouxe-nos o advento do documento digital. Nem todo registro de informações que utiliza a eletrônica para gravação e reprodução faz uso da tecnologia digital, ou seja, nem todo documento eletrônico é digital, como, por exemplo, o disco em vinil. De qualquer forma, os documentos digitais vêm, cada vez mais, assumindo uma posição de destaque em vários aspectos da vida contemporânea: é o caso da *fotografia digital* ou dos *arquivos de imagens* gerados no processo de digitalização de documentos em suporte papel[i]. As disciplinas que utilizam documentos como matéria-prima de trabalho não poderiam deixar de ser afetadas pela presença do documento digital. É o caso da História, da Biblioteconomia e da Arquivologia, entre tantas outras.

Um dos problemas mais instigantes que se apresenta em função da existência do documento digital é sua preservação. Aqui cabe uma distinção entre os termos *preservação*, *conservação* e *restauração*. Segundo Muñoz Viñaz, o termo *conservação* pode ser entendido num sentido restrito em oposição à idéia de *restauração*, ou seja, atividades para manter (*keep*) o original ou, num sentido mais amplo, significando a soma dessa primeira idéia e outras atividades possíveis relacionadas. O mesmo autor acredita que há uma confusão terminológica:

> A confusão surge porque nas línguas latinas como o italiano, espanhol ou francês, 'conservation' num sentido mais amplo traduz-se por 'restauro' (italiano), 'restauración' (espanhol') ou 'restauration' (Francês), de maneira que as traduções dessas línguas para o inglês e vice-versa são freqüentemente imprecisas. As coisas ficam ainda piores porque alguns autores e organizações usam diferentes sinônimos para 'conservation' num sentido amplo, como o termo 'preservation' e até mesmo 'restoration' (MUÑOZ VIÑAZ, 2005, p. 14, tradução nossa).

Neste trabalho, utilizaremos o termo ***preservação***, preterindo o termo *conservação*, seguindo, assim, uma tendência entre os autores que publicam sobre preservação digital. O sentido do conceito de preservação que empregamos aqui é próximo ao que Muñoz Viñaz chama de sentido amplo do termo '*conservation*', ou seja, diversas atividades que podem ser feitas para assegurar a **integridade** e o **acesso** aos documentos pelo maior prazo possível, idealmente para sempre. Uma excelente definição de preservação de documentos digitais foi exposta por Conway: "Preservação [*preservation*] é a aquisição, organização e distribuição de recursos a fim de que venham a impedir posterior deterioração ou renovar a possibilidade de utilização de um seleto grupo de materiais" (CONWAY, 2001, p. 14).

Um pesquisador atento ao problema da **preservação** de documentos digitais pode se preocupar com diferentes expectativas de vida. Diferentemente de documentos em papel de boa qualidade ou do microfilme de guarda permanente, documentos digitais podem se tornar imprestáveis em uma década ou menos, se os devidos cuidados não forem aplicados: "Durante o século XX, a permanência, durabilidade e a resistência dos mais recentes meios de registro, com exceção do microfilme, continuaram a declinar" (SEBERA, 1990, apud CONWAY, 2001, p. 13).

Percebe-se, então, que mesmo documentos digitais que precisam ser mantidos por algumas décadas por motivos administrativos, contábeis ou fiscais, podem não durar o suficiente para cumprir sua função original. No entanto, o problema certamente é bem mais sério quando nos referimos aos documentos digitais que necessitam ser mantidos por séculos à frente, tanto quanto for possível, para as gerações futuras. Esses documentos compõem um legado cultural e histórico para a humanidade. Nesse projeto de pesquisa, nossa atenção se

volta para a preservação dos documentos digitais de cunho histórico e cultural e que, por isso, necessitam de guarda permanente.

Há que se distinguir, também, no que diz respeito aos documentos digitais, por um lado, os aspectos relacionados à preservação dos suportes físicos utilizados, como CDs e fitas magnéticas; e, por outro lado, o próprio conteúdo informacional existente nos documentos. Tomemos como ilustração uma reportagem fotográfica histórica que utiliza a tecnologia digital: *as filmagens no atentado de 11 de setembro nos Estados Unidos*. Aquelas imagens foram gravadas e (re)gravadas em inúmeros suportes, CDs, discos em servidores de rede na internet, fitas magnéticas, etc. Cada um desses suportes documentais tem suas próprias necessidades de preservação, as quais, aliás, são muito relevantes, pois sua vida útil costuma ser bem pequena, sem mencionar o fato de que são suportes físicos muito mais frágeis que o papel, por exemplo. Portanto, um mesmo conteúdo informacional pode estar presente em diferentes suportes físicos, concomitantemente ou não. Além disso, esse conteúdo informacional – imagens, no exemplo citado – também apresenta seus próprios problemas do ponto de vista da preservação por longos períodos.

No projeto de pesquisa ora em desenvolvimento, nosso escopo compreende os objetos digitais que codificam conteúdos como imagens em movimento ou fixas, texto, som ou uma combinação desses elementos. Não estamos preocupados, portanto, com a preservação de suportes físicos utilizados nos documentos digitais. Por outro lado, indiretamente, nosso trabalho afeta a preservação de documentos em suportes tradicionais, aqueles nos quais não é possível uma separação entre conteúdo e suporte físico, como livros em papel, mapas, etc.

A intersecção entre a preservação de documentos em suportes tradicionais e a preservação de objetos digitais ocorre em função do processo de digitalização. Em si, esse processo tem sido utilizado como vetor da preservação, pois os objetos digitais gerados atualmente podem conter uma alta fidelidade aos originais, o que permite poupar o acesso direto e o manuseio dos originais. Além disso, **caso se obtenha êxito na preservação desses objetos digitais**, é possível que esses persistam mesmo após a inevitável degradação física dos suportes utilizados nos documentos tradicionais, como o papel comum, os diferentes tipos de papel fotográfico, a película cinematográfica, etc. Sobre o processo de digitalização e os cuidados com os objetos digitais gerados, Paul Conway observa que:

> Imagens digitais estão se tornando realmente comuns em bibliotecas e arquivos. A qualidade dos produtos de imagem digital pode ser espetacular. Há pouca dúvida de que a qualidade irá melhorar acompanhando a maturidade da tecnologia. Organizações estão reorganizando orçamentos, arrecadando dinheiro e antecipando receitas para fazer os projetos digitais acontecerem. Pode alguma instituição – bibliotecas, arquivos, sociedades históricas ou museus – arcar com o desperdício desse investimento? Sem um esforço sério que assegure o acesso por longos períodos dos arquivos digitais de imagens, porém, o risco de perdas é tremendo (CONWAY, 2000, tradução nossa).

Um outro aspecto que também relaciona a preservação de objetos digitais aos documentos tradicionais é a possibilidade de restauração dos últimos, tomando-se como referencial a imagem dos primeiros. Sobre isso:

> Considerar um repositório digital de artefatos culturais não apenas como uma ferramenta educacional e de história da arte, mas também como uma poderosa ferramenta de restauração, implica que, além das informações visuais (imagens, raios-x, etc.) e informações textuais/metadados simples, uma abundante quantidade de dados para pesquisa/restauração deveriam ser armazenados no repositório (DELOS-NSF Working Group, 2002, p. 4).

Os objetos digitais aos quais nos referimos nesse trabalho são constituídos por dígitos binários. Qualquer objeto digital, em última análise, independentemente do tipo de conteúdo (texto, som, imagem, etc.) ou tipo de suporte físico onde será gravado (disco rígido, fita magnética, etc.) será sempre composto por um conjunto de números binários. Esse conjunto somente é legível através de *hardware* e *software* apropriados. Mesmo esses dois elementos só podem interpretar esses dígitos através de um enunciado que "explica" o significado desses *bits*. Por exemplo, é preciso indicar se um trecho de *bits* corresponde à data de gravação do arquivo, o tipo de arquivo ou parte do texto (quando se tratar de um arquivo de texto) ou parte do som (caso se trate de um arquivo de som). Esse enunciado é conhecido como **Especificação do Formato de Arquivo** ou, simplesmente, **Formatos de Arquivo** (*File Formats*).

Não tentaremos desenvolver um aprofundamento técnico sobre o que são formatos de arquivo e suas especificações, pois isso foge ao escopo dessa introdução. No entanto, nossa pesquisa focaliza justamente o conceito técnico de formato de arquivo: **identificando** as características mais adequadas que subsidiem a escolha de determinado formato de arquivo para a preservação de guarda permanente e efetuando um **levantamento** dos formatos de arquivo efetivamente em uso, dessa forma **diagnosticando** o quadro atual no que diz respeito aos efeitos na preservação de documentos digitais para as gerações futuras.

Acreditamos que a melhor justificativa para esse trabalho reside no próprio papel das bibliotecas contemporâneas, juntamente com a própria universidade, papel esse que Donald Waters assim define:

> Eu afirmaria que a missão da universidade e da **biblioteca** é produzir cidadãos cultos. A função ampla da universidade dando suporte a essa missão, incluindo a preservação do conhecimento, está sendo mantida, mas os meios da comunicação acadêmica pela qual a universidade efetua essas várias funções estão hoje em mutação. A comunidade acadêmica precisa se ajustar às mudanças nos meios de comunicação e porque os **programas de preservação** são, por definição, o principal mecanismo para renovar os ativos da universidade e da **biblioteca**, eles podem e devem ajudar nos necessários ajustes (WATERS, 1998, p. 100, grifos e tradução nossa).

Em consonância com essa linha de pensamento, as grandes bibliotecas do planeta vêm desenvolvendo programas voltados para a preservação de documentos digitais, mais especificamente preocupados também com o problema dos formatos de arquivo. A *British Library* mantém um programa de preservação digital com vários projetos, muitos levados a cabo com outras instituições[ii]. Aliás, considerando o custo de pesquisa em preservação digital, além de outros fatores, tem-se defendido o trabalho em cooperação. Nesse sentido:

> O fato de que a preservação digital é cara, os fundos são escassos e as responsabilidades são difusas sugere que as atividades de preservação digital se beneficiam da cooperação. Cooperação pode incrementar a capacidade de produtividade de um suprimento limitado de fundos de preservação digital através do compartilhamento de recursos, eliminado redundâncias e explorando a economia de escala. (LAVOIE, DEMPSEY, 2004, tradução nossa).

Nos Estados Unidos, a *Library of Congress* também mantém diversos projetos especificamente sobre preservação digital: "Em muitos casos, materiais digitais são considerados mais frágeis que seus correspondentes físicos. Os arquivos em si podem facilmente ser destruídos ou armazenados em um formato que se torne obsoleto"[iii].

Entre tantas instituições de renome mundial, a biblioteca da Universidade de Harvard mantém um programa específico para tratar do problema dos formatos de arquivo. O projeto JHOVE[iv] tem como objetivo propiciar hoje para as gerações futuras as funções de validação, identificação e caracterização de formatos de arquivo (*representation format*): "As ações de identificação, validação e caracterização são freqüentemente necessárias durante a operação de rotina de repositórios digitais e para a preservação digital"[v].

2.0 - Resultados obtidos

2.1 - O que é um Formato de Arquivo

Sem dúvida, essa parte conceitual é a mais importante em nosso trabalho; pode-se dizer que se tratará da alma da dissertação. É essa base conceitual que norteia toda a coleta de dados que estamos implementando. Devemos aqui responder a pergunta fundamental: *O que são Formatos de Arquivo?*

Esse conceito parece padecer do mesmo problema que o conceito de documento. Esse é um conceito prosaico e com o qual quase todas as pessoas lidam em seu dia-a-dia. E, pelo mesmo motivo, ou seja, por ser largamente utilizado, apresenta vários sentidos, dependendo de quem o interpreta e utiliza. O resultado é um conceito "fácil", todos sabem o que é, todos podem dizer o que é e, conseqüentemente, fica cada vez mais difícil defini-lo com precisão.

No caso do conceito de documento, no âmbito dos pesquisadores da área de Documentação e Ciência da Informação, sabemos o quanto é difícil defini-lo precisamente.

Com o objetivo, então, de definir com a maior precisão e clareza possível o conceito de Formato de Arquivo, iniciaremos esta parte trazendo algumas definições presentes em outros trabalhos de pesquisa. Antes, porém, vamos trazer à luz alguns conceitos ainda mais fundamentais.

2.1.1 - Digital e analógico

O uso do termo digital é bastante novo na humanidade, pelo menos na acepção que aqui nos interessa, ou seja, a que tem sido utilizada em tecnologia eletrônica e informática. Um aspecto fundamental desse termo se refere a uma nova maneira de registrar e representar informações.

Os primeiros artefatos eletrônicos que o homem criou utilizavam exclusivamente o que agora chamamos de tecnologias analógicas, contrapondo-se às atuais tecnologias digitais. Alto-falantes utilizados em qualquer equipamento de som, como as caixas de som do computador, são um bom exemplo de tecnologia analógica. O som produzido por esses equipamentos é o resultado do movimento mecânico de eletroímãs, as características sonoras como os graves e agudos e a altura do som são o resultado de milhares de movimentos mais ou menos intensos; ocorre uma miríade de movimentos.

Atualmente, apesar de ainda utilizarmos a tecnologia analógica em muitos equipamentos, como no exemplo acima, a maioria dos circuitos internos de qualquer equipamento eletrônico processa sinais no modo digital. Em oposição à miríade de opções exemplificada acima, há um número finito de opções: zeros e uns. Apesar do exemplo dado no universo dos equipamentos sonoros, sem dúvida, a maior aplicabilidade da tecnologia digital está no âmbito da informática: armazenar e processar informações representadas pelos números zero e um.

Um estudo aprofundado dessa tecnologia tomaria muitas e muitas páginas, mas o que nos interessa é o aspecto da codificação binária.

2.1.2 - Codificação binária

O princípio fundamental do uso de tecnologia digital no universo da informática é o de converter as informações utilizadas na linguagem humana – como o nosso sistema de escrita e numeração – em códigos formados por grupos de números binários: somente o número zero e o número um. Naturalmente, o número de dígitos necessários para representar essas informações dependerá da complexidade das informações a serem representadas.

Assim, com 3 dígitos binários podemos representar $2^3 = 8$ códigos:

000
001
010
011
100
101
110
111

Os computadores atuais – além de outros dispositivos digitais – trabalham, atualmente, com códigos de 64 dígitos ou mais. Essa quantidade de códigos permite armazenar uma grande quantidade de informações. Muito além dos caracteres de nossa linguagem (em qualquer idioma), é possível representar as cores utilizadas numa imagem (em cada minúsculo ponto), os sons de uma música ou a fala humana, isso sem mencionar os códigos internos, que possuem significado somente para os circuitos, como os comandos dos microprocessadores ou endereços de memória.

2.1.3 - Definições

Vamos agora trazer à luz o conceito de formato de arquivo e relacioná-lo com a representação no universo digital.

Num relatório elaborado no âmbito do projeto *The Representation and Rendering Project*, da Universidade de Leeds, no Reino Unido, encontramos a seguinte definição para formato de arquivo:

> Em seu nível mais baixo, objetos digitais são seqüências de zeros e uns que representam dados codificados. Diferentes Formatos de Arquivo especificam como esses códigos representam o conteúdo intelectual criado por um autor de um objeto digital. (UNIVERSITY OF LEEDS, [s.d], p. 4, tradução nossa).

A definição chama a atenção para o fato de que um formato de arquivo qualquer especifica como um determinado conteúdo está estruturado.

O termo técnico associado ao "como" da definição anterior chama-se *especificação*. Sobre esse termo: "Uma definição completa de formato de arquivo tem de incluir o conceito de especificação (*specification*), o qual em si pode ser definido como os requisitos organizacionais de um arquivo" (SHEPARD; MacCARN, [s.d], p. 6, tradução nossa).

Os "*requisitos organizacionais de um arquivo*" se referem à estrutura em que os códigos digitais estão organizados para cada tipo de arquivo (formatos de arquivo). Essa estrutura extrapola em muito os códigos utilizados para representar o conteúdo de um arquivo, seja ele texto, imagem, som ou outro qualquer. Além do conteúdo, muitas outras informações

são necessárias. Tomemos como exemplo um arquivo de texto simples contendo uma pequena receita. Na tela de um aplicativo editor de texto ele seria visualizado como na figura 1:

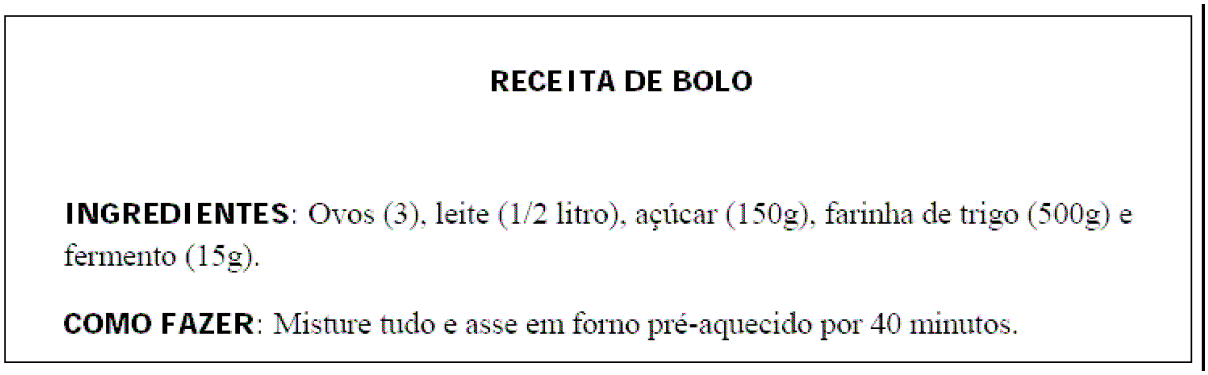

Figura 1 – Arquivo com texto visualizado através de um editor de textos.

Que informações deveriam ser gravadas no arquivo correspondente ao conteúdo do texto acima? Em primeiro lugar, o próprio texto. Ou seja, os códigos binários que correspondem aos caracteres utilizados acima. Notemos também que foram utilizados caracteres com as fontes *Tahoma* e *Times New Romam*. Além disso, algumas palavras estão em negrito. Há também informações sobre os espaços entre linhas e entre caracteres, margens, etc. Essas informações todas se referem ainda ao conteúdo visível do texto. Porém, um arquivo real necessita também *metadados* mínimos como a data de criação do arquivo, o tamanho desse arquivo em *bytes*, o *software* utilizado para a sua criação, etc.

Quando lidamos com arquivos de imagens fixas, som ou imagem em movimento, o grau de complexidade aumenta consideravelmente.

Uma especificação para um formato de arquivo X nada mais é senão a determinação de quais informações (conteúdo, *metadados* e outros) e ordem seqüencial (ou não) de gravação no arquivo físico composto de códigos binários.

Infelizmente, a primeira coisa a reconhecer é o quanto uma especificação de formato de arquivo não é simples, desde os menos complexos arquivos de texto até formatos de arquivo específicos para imagens em movimento.

Vamos fazer uma pequena análise numa especificação real de formatos de arquivo com o objetivo de compreender ainda melhor esse conceito tão importante. Escolhemos uma especificação menos complexa tomando como parâmetro o poder de processamento e recursos do aplicativo que gera o arquivo nessa especificação: o aplicativo *WRITE*, um editor de texto da empresa *Microsoft*.

No início da primeira página, há uma orientação sobre características básicas dessa especificação. Sabemos que esse tipo de arquivo contém, além do conteúdo propriamente dito, texto e figuras e formatação.

O primeiro tópico abordado tem o título de *File Header* (cabeçalho do arquivo), que descreve o conteúdo do arquivo; por exemplo, no cabeçalho está registrado o comprimento do arquivo (*length of the file*). Logo abaixo temos acesso a uma tabela com as *Word* (palavras), *Name* (nomes das palavras) e suas respectivas descrições. Cada *Word* corresponde a 16 *bits*[vi]. A primeira *word* (*wIdent*) parece ser utilizada para identificar o arquivo; normalmente teria o número 0137061 (em linguagem octal), que corresponde a 1011111000110010 (em linguagem binária[vii]).

Ainda na primeira página da especificação, ao final, encontramos um tópico com o título *Text* (texto). Nesse tópico ficamos sabendo que o texto num arquivo desse tipo inicia a partir da *word* 64 na página 1. Mais adiante sabemos que os caracteres ASCII[viii] de números

13 e 10 têm uso especializado e correspondem respectivamente ao comando para retorno de cada linha num parágrafo (*carriage return*) e avanço para uma próxima linha (*linefeed*).

Na seqüência temos ainda mais 6 páginas e tópicos relacionados às *Pictures* (figuras) eventualmente utilizadas no arquivo, *Formatting* (formatação), *Characters and Paragraphs* (caracteres e parágrafos), *Sections* (seções num mesmo documento) e informações sobre as fontes de caracteres utilizadas (*Font Table*). Facilmente percebemos que se trata de informações bastante especializadas, compreensíveis e úteis para iniciados em linguagens de programação e Ciência da Computação. Nosso objetivo, aqui, é apenas exemplificar uma especificação real de formato de arquivo.

2.2 – Tipos de Formatos de Arquivo

Existe hoje uma grande quantidade de especificações técnicas para uma infinidade de formatos de arquivo diferentes. Muitas das especificações atualmente em uso evoluíram a partir de versões antigas de aplicativos hoje descontinuados. Além disso, *software* novo é criado diariamente; conseqüentemente, novas especificações de formatos também. A grande explosão de novos formatos de arquivo ocorreu com o surgimento da microinformática e os computadores pessoais; mas, antes desse período – últimas décadas do século XX – já existiam no mundo dos *mainframes*[ix]. Segundo Kientzle, "Sistemas operacionais para *mainframes* tratam um arquivo como um repositório de base de dados. Cada item nessa base de dados é um *record*[x] e, dessa forma, *mainframes* tratam arquivos como uma coleção de *records*[x]" (KIENTZLE, 1995, p. 358, tradução nossa).

2.2.1 – Classificação de Formatos de Arquivo

Uma primeira classificação de formatos de arquivo pode ser feita com base no tipo de *software* utilizado para gerar os arquivos que serão gravados em algum tipo de mídia de acordo com a **especificação** do formato. No exemplo que utilizamos antes em 1.2, o formato de arquivo ***Write*** seria do tipo Texto, pois é gerado através de um aplicativo para edição de **textos**. Essa classificação é problemática, no entanto, pois, em geral, podemos falar em aplicativos que geram predominantemente texto, imagens fixas, sons, etc. Isso ocorre mesmo em formatos de arquivo aparentemente exclusivos para certos conteúdos. Um exemplo é o formato de arquivo MP3 feito especialmente para registro de sons em geral. Ocorre que é possível incorporar ao arquivo no formato MP3 legendas **textuais** para as músicas gravadas. Um outro exemplo nesse sentido se refere ao formato GIF, projetado para imagens fixas, apesar de existir o chamado GIF animado, que pode incorporar imagens em movimento. Assim, em geral, pode-se falar de formatos de arquivo para conteúdos predominantemente em determinado conteúdo. Para isso, consulte a tabela abaixo:

Tipo predominante de conteúdo	Exemplos de Formatos de Arquivo
Texto	RTF, OpenOffice, ODF, DOC, AmiPro e outros
Imagens fixas	BMP, EXIF, GIF, JPG, TIFF e outros
Imagens em 3D	CAD, BIFF, X4D e outros
Sonoro	MEU, KAR, MP3, MP4 e outros
Imagens em movimento	AVI, MOV, MPEG, SWF e outros

Tabela 1 – Classificação de formatos de arquivo pelo conteúdo.

Note que na tabela acima os exemplos de formatos de arquivo são nomeados pela extensão do nome do arquivo em ambientes de computadores pessoais (*Windows*, *MacOS* e outros); discutiremos sobre extensões na parte sobre identificação de formatos de arquivo. A tabela acima não é exaustiva mas apenas ilustrativa; no sítio *Wotsit.org* (http://www.wotsit.org) é possível consultar uma relação bem mais completa de especificações de formatos.

2.2.2 – Versões de Formatos de Arquivo

Nesse ponto é essencial chamar a atenção para um detalhe técnico extremamente importante: formatos de arquivo possuem, geralmente, diferentes versões. Desde a primeira versão de um *software*, digamos, um editor de textos, várias modificações e aperfeiçoamentos são implementados. Por exemplo, em editor de texto pode não permitir o uso de imagens junto ao documento textual; mas, a partir de uma nova versão, esse recurso passa a ser possível. Assim, haverá modificações na especificação original do formato de arquivo para que seja possível armazenar imagens nos arquivos. Algumas novas versões de um mesmo formato de arquivo podem ser consideravelmente diferentes da versão anterior, além da própria freqüência com que surgem novos formatos:

> Versões de formatos de arquivo tendem a ter vida curta em função de interesses comerciais dos desenvolvedores de software. As aplicações de software geralmente não permitem facilidades de importação para todas as versões anteriores de formatos de arquivo (UNIVERSITY OF LEEDS, [s.d], p. 4, tradução nossa).

2.3 – Características adequadas para preservação

Uma das conclusões mais importantes nesse trabalho se refere à definição de quais são as características mais relevantes que um determinado formato de arquivo deve possuir para que seja considerado como indicado para guarda por longos períodos; ou seja, quais são as características que aumentam as chances de que um arquivo continue garantindo acesso ao seu conteúdo.

Uma das características específicas que tem sido apontada como extremamente importante para fins de preservação por longos períodos é o acesso público à especificação do formato de arquivo, o que também é conhecido como uso de *formatos abertos* de arquivo. Como já abordamos anteriormente, todo formato de arquivo possui uma especificação, mas essa não é, necessariamente, de acesso público. As vantagens por trás da utilização de formatos de arquivo abertos se evidenciam ao se prever a necessidade de desenvolvimento de *softwares* para leitura desses documentos no futuro. O trabalho de desenvolvimento pode ser até mesmo inviável caso não se conheça os detalhes técnicos de determinado formato de arquivo. O TIFF (*Tagged Image File Format*) é um bom exemplo de especificação de formato de arquivo aberto; na verdade, em função de sua popularidade, existem vários grupos de discussão sobre esse formato, como o *LibTiff Mailing List*[xi].

No extremo oposto aos formatos de arquivo abertos, encontramos os formatos proprietários, como aqueles da família *Microsoft Office 96*: .doc, .xls e outros. Empresas como a própria *Microsoft* têm sentido a pressão pela abertura de suas especificações; nesse sentido, têm surgido possíveis soluções como o uso da linguagem XML na gravação dos arquivos.

Uma outra característica importante é a padronização de formatos de arquivo. O fato de um formato ter sua especificação aberta não significa necessariamente padronização em sua especificação. O efeito de uma especificação aberta pode até mesmo ser danosa para a padronização, na medida em que pequenas "melhorias" podem ser incorporadas por diversas empresas e desenvolvedores. O problema é que essas "melhorias" podem não ser suficientemente documentadas e divulgadas para todos. A criação de normas oficiais para formatos de arquivo traz ainda a vantagem de impor uma especificação fixa, que pode até mesmo incluir modificações a partir da versão original, mas, nesse caso, sendo documentadas adequadamente. Um dos melhores exemplos de formatos de arquivo padronizados é o formato PDF/A; sobre isso, "O formato promete ser largamente aplicável na criação e distribuição de documentos, registrando evidências de transações, buscando e recuperando e muitos outros usos comuns." (LeFURGY, 2003, tradução nossa).

3.0 – Considerações Finais

Conhecer o próprio conceito de formato de arquivo, incluindo os conceitos de especificação, versão e características adequadas para a preservação por longos períodos é uma condição *sine qua non* para o sucesso num programa de preservação digital.

É preciso frisar que o uso de formatos de arquivos adequados para a preservação não precisa ocorrer desde a criação dos documentos digitais, apesar disso ser desejável. Muitos documentos digitais serão criados de acordo com o *software* correspondente a um determinado formato de arquivo, mesmo que esse não seja adequado para a preservação por longos períodos. Será difícil convencer uma instituição a utilizar somente determinados formatos de arquivo com base no argumento da preservação do documento digital. Até porque a imensa maioria dos documentos não necessita de guarda permanente. Estima-se que entre 90 e 95% dos documentos de uma organização têm como destinação o descarte e não a guarda permanente. Por outro lado, aquela pequena fatia de documentos que deve ser preservada para a posteridade pode ser migrada para um formato com as características adequadas para a preservação, por exemplo, especificações abertas.

Referências

CONWAY, P. From analog to digital: extending the preservation tool kit. In: DeWITT, D. L. **Going digital**: strategies for access, preservation, and conversion of collections to a digital format. London: Haworth Press. p. 65-79, 1998.

_____. **Preservação no universo digital**. 2ª ed. Rio de Janeiro: Projeto Conservação Preventiva em Bibliotecas e Arquivos: Arquivo Nacional, 2001.

_____. Overview: rationale for digitization and preservation. In: SITTS, Maxine K. **Handbook for digital projects**: a management tool for preservation and access. Massachuttes: Northeast Document Conservation Center, 2000. Acesso em 15/4/2008. Disponível em: < http://nedcc.org/oldnedccsite/digital/dighome.htm >

DELOS-NSF Working Group on Digital Imagery for Significant Cultural and Historical Materials. Edited by Ching-chih Chen and Kevin Kiernan. December 2002. Acesso em 15/4/2008. Disponível em < http://dli2.nsf.gov/internationalprojects/working_group_reports/digital_imagery.html >.

KIENTZLE, Tim. **Internet file formats**. Arizona: Coriolis Group, 1995.

LeFURGY, William G. PDF/A: Developing a file format for long-term preservation. **RLG News**. Nova York, v. 7, n. 6, 2003. Disponível em: http://www.rlg.org. Acesso em: 10/11/2005.

LAVOIE, B.; DEMPSEY, L. Thirteen ways of looking at... digital preservation. In: **D-Lib Magazine**, v. 10, n. 7/8, 2004. Acesso em 15/4/2008. Disponível em < http://www.dlib.org/dlib/july04/lavoie/07lavoie.html >.

MUÑOZ VIÑAS, Salvador. **Contemporary theory of conservation**. Reino Unido: Elsevier, 2005.

SHEPARD, Thom; MacCARN, Dave. **The universal preservation format**: a recommended practice for archiving media and electronic records. Boston, [s.d]. Disponível em http://info.wgbh.org/upf/. Acesso em 22/3/2008.

UNIVERSITY OF LEEDS. **Survey and assessment of sources of information on file formats and software documentation**. The representation and rendering project. Reino Unido, [s.d]. 48 p. Disponível em http://www.leeds.ac.uk/reprend. Acesso em 22/3/2008.

WATERS, Donald. Transforming libraries through digital preservation. In: **Going Digital**: strategies for access, preservation, and conversion of collections to a digital format. New York: The Haworth Press, 1998.

Notas

[i] Há que se fazer uma distinção entre documentos digitais nascidos digitais e aqueles gerados a partir da digitalização de documentos tradicionais. A digitalização, atualmente, é um processo que se aplica para praticamente todos os gêneros documentais: imagem, som e texto.

[ii] Pode-se conhecer melhor os programas de preservação digital da *Britsh Library* em < http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/index.html >

[iii] Acessado em 15/04/2008. Disponível no sítio da *Library of Congress*: < http://www.digitalpreservation.gov/you/digitalmemories.html >.

[iv] JHOVE, JSTOR/*Harvard Object Validation Environment*, "Format-Specific Digital Object Validation," 2004. Disponível em < http://hul.harvard.edu/jhove/index.html >.

[v] [v] Disponível em < http://hul.harvard.edu/jhove/index.html >.

[vi] Uma *Word* de 16 *bits* é uma convenção utilizada em linguagens de programação e significa um número com 16 dígitos binários.

[vii] As representações em linguagem octal, binária ou outras como a hexadecimal e decimal (a utilizada por nós no dia-a-dia) são apenas maneiras diferentes de representar quantidades numéricas e cada uma é mais apropriada para determinado uso.

[viii] ASCII, lê-se *ásqui 2* e significa *American Standard Code for Interchange of Information*. Trata-se de uma tabela com códigos binários e seus correspondentes a caracteres comuns, especiais ou comandos específicos.

[ix] O termo *mainframe* é utilizado para designar computadores de grande porte, utilizados apenas por grandes corporações na era anterior à microinformática. É curioso notar que, na verdade, possuíam poder de processamento inferior aos computadores pessoais atualmente em uso.

[x] Um *record* ou registro numa base de dados corresponde a cada grupo de campos. Por exemplo, os campos nome, idade e endereço exigirão tantos registros quantos forem os nomes da relação de pessoas numa organização.

[xi] Acesse em http://www.asmail.be/msg0054995370.html.