PRESERVAÇÃO DE DOCUMENTOS ELETRÔNICOS: O PAPEL DOS FORMATOS DE ARQUIVO

Ernesto Carlos Bodê

INTRODUÇÃO

Esta pesquisa teve início a partir da pura necessidade de buscar respostas e soluções para problemas de preservação relativos a acervos documentais compostos por documentos em suportes não tradicionais, como CD´s e Fitas magnéticas, além do tipo de suportes, o conteúdo destes documentos parecia e de fato possuem características específicas, como a necessidade de software para lê-los (dependência de software). A produção brasileira sobre o assunto contém alguns trabalhos. Em 1999, publicado pelo *Council on Library and Information Resources* estadunidense, surge um relatório (BECK, 1999) de Ingrid Beck que "conta uma história" sobre um projeto de tradução de textos para nosso vernáculo, a publicação, no que diz respeito à preservação no universo digital menciona a tradução de um excelente texto de Paul Conway de 1996, *Preservation in a Digital World.* Em 2001, Marcelo Leone Sant'Anna publica "Os Desafios da Preservação de Documentos Públicos Digitais" (SANT'ANNA, 2001). No início de 2004 temos o trabalho "A preservação digital e o modelo de referência Open Archival Information System (OAIS)" de Kátia P. Thomaz e Antônio J. Soares 2004 (THOMAZ; SOARES, 2004). Em meados de 2004 Miguel A. Arellano, publica uma excelente "pesquisa bibliográfica", "Preservação de Documentos Digitais" (ARELLANO, 2004).

No entanto, como fica claro até mesmo pela bibliografia dos trabalhos acima, a grande fonte de informações sobre os diversos aspectos da preservação de documentos digitais são os grandes projetos internacionais como o *Research Library Group* (RLG), o *Council on Library and Information Resources* (CLIR) e diversos outros grupos em vários continentes além da América do Norte, principalmente Europa e Oceania. Estes grupos e projetos se constituíram em nossa principal fonte de pesquisa, com excelentes relatórios, artigos, *surveys* e etc.

Inicialmente, nos deparamos com uma grande gama de problemas e possíveis soluções, paulatinamente, fomos amadurecimento nossa visão e passamos a enfocar um ponto

específico, a nosso ver fundamental como alicerce para várias outras soluções, trata-se do tema Formatos de Arquivo. Neste artigo procuraremos definir o que é e qual o papel desempenhado pelos objetos 'formatos de arquivo' no que cabe às atividades de preservação de documentos.

Entendo que há alicerces estabelecidos, como resultado de pesquisas e discussões na área de conservação, preservação e restauração de bens culturais, e este alicerces não podem ou devem ser ignorados em trabalhos que envolvam este novo objeto de acervo que é o documento eletrônico. Segundo Salvador Muñoz Viñaz (MUÑOZ-VIÑAZ, 2005), o termo conservação pode se referir a dois sentidos, o primeiro deles em oposição a *restauração*, um sentido mais restrito de atividades e o segundo como a soma das atividades do primeiro sentido mais *restauração* e outras atividades correlatas, tais como pesquisa histórica e apoio administrativo. Ainda segundo Muñoz, os problemas de tradução entre línguas latinas e outros ramos linguísticos são responsáveis pelas imprecisões como o uso do termo preservação como conservação e vice-versa. Segundo 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), ou seja, conservação num sentido amplo. Neste trabalho, utilizaremos o termo preservação, pois em geral, estamos nos referindo a atividades mais amplas que também incluem a conservação dos documentos (atividades para mantê-los em uso).

Agora, abordaremos mais detalhadamente o documento que utiliza eletrônica e codificação digital. Para que possamos compreender melhor as características e limites destes documentos, elaboraremos uma breve evolução histórica do documento de maneira geral, até chegarmos na atualidade com as características específicas de nosso objeto.

UM BREVE HISTÓRICO DO DOCUMENTO

Nos pareceu óbvio a existência de uma forte ligação entre a evolução do documento e a própria evolução dos sistemas de escrita. Desta forma, seguiremos inicialmente esta trilha, tentando dar foco também nos suportes utilizados neste documentos, o que será importante mais adiante como comparação entre épocas.

Segundo Steven R. Fischer (2003), até que a humanidade obtivesse sistemas de escrita completos como os atuais, ela fez uso de símbolos gráficos e mnemônicos de vários tipos para armazenar informações, sobre um dos mais antigos artefatos encontrados, escreveu:

> Artefatos desenterrados em Bilzingsleben, Alemanha, datas de pelo menos, 412.000 anos atrás [...] foram interpretados por seus descobridores como entalhes intencionais (algum tipo de símbolos gráficos). É evidente que os entalhes são marcas; o que significam e se significam algo, não está claro. (FISCHER, 2003, p. 16, tradução nossa)

Ainda segundo Fischer (2003), a humanidade utilizou então sistemas pictográficos (como as representações em cavernas), em um segundo momento passou a utilizar símbolos gráficos para representar objetos reais como mulher, animais e etc., até o grande salto da fonetização, quando um símbolo gráfico representa um som da linguagem local. Tal invenção surgiu na Mesopotâmia entre 6.000 e 5.700 anos atrás.

Como suportes para registros de informações a humanidade utilizou os mais diversos materiais, segundo Dard Hunter (1978), foram utilizados a madeira, metais, pedras, troncos, tecidos, o papiro (Cyperus papyrus), pergaminho e finalmente o papel.

O uso de argila em tábuas é particularmente importante, pois ao que parece o primeiro sistema completo de escrita (por volta de 2.500 AC) utilizou este material como suporte. Inicialmente, com o desenvolvimento da escrita cuneiforme (que foi utilizada por vários povos) e cujo princípio, símbolos representando sons, foi a base dos hieróglifos egípcios, que deu origem ao primeiro alfabeto da humanidade.

No Egito, o uso do papiro rivalizou com o uso de tábuas de argila. Na verdade os egípcios desenvolveram diferentes sistemas de escrita para diferentes aplicações, rituais, contabilidade e etc., para cada aplicação havia um sistema de escrita e suportes específicos como paredes, ouro e etc.

O uso de pergaminho também foi um fato importante para o registro de informações, "O rei de Pérgamo (197-159 AC) normalmente recebe os créditos pela invenção e acredita-se que esteja relacionada com o desejo de produzir um material de escrita que rivalizasse com o papiro egípcio". (HUNTER, 1978, p. 12, tradução nossa)

Finalmente, a invenção do papel possibilitou um grande salto na produção de documentos, pois tratava-se de um material de fácil fabricação e menor custo, além da qualidade em relação a outros suportes. A data normalmente atribuída à invenção do papel é a de 105 DC, na China (HUNTER, 1978).

Vários outros suportes foram utilizados para o registro de documentos e em determinados períodos históricos alguns competiram entre si, como o papel e o pergaminho. O tipo de papel próximo do que é utilizado hoje só existiu a partir do século XIX (DOCTORS, 1999).

Temos então, até o século XIX, uma produção documental, registrada basicamente em papiro, pergaminho e papel, documentos com conteúdo textual, de diferentes

naturezas, de inventários de bens a literatura e filosofia. Em meados do século XIX surge uma invenção que acrescenta uma nova diversidade aos acervos documentais, trata-se da fotografia, "A invenção da fotografia foi anunciada oficialmente em 19 de agosto de 1939, pelo francês Louis Jacques M. Daguerre (1787-1851), sob a forma do daguerreótipo" (SMIT; GONÇALVES, 2005, p. 9). Esta invenção passaria por um processo de evolução tecnológica que culminaria, no final do século XX com a advento da fotografia digital, a qual, por si só, tornou-se uma nova revolução. Também com tecnologia bastante próxima dos registros fotográficos, apesar da aplicação diferente, encontramos também o microfilme como meio para registro documental, ainda hoje bastante utilizado.

No final do século XIX, vários inventos para registro do som culminaram no início do século XX com os discos de áudio, primeiro e logo depois, o uso também de fitas magnéticas. Estas últimas, após um período de evolução, também utilizadas para gravação de vídeo (os primeiros programas televisivos gravados). No final do século XX surgiriam os discos do tipo CD (Compact Discs), inicialmente para gravações de áudio (CD-ROM) e depois surgindo os modelos específicos para vídeo (DVD's). A miríade de opções de CD e DVD - além do tipo ROM, os outros formatos mais conhecidos são os do tipo R (*Recordable*) e RW (*Rewritable*) - seria logo aproveitada juntamente com a tecnologia de computadores.

O próximo grande passo seria dado pelo uso de computadores pela humanidade,

> Os primeiros computadores modernos apareceram na década de 40; embora tenha havido muitas contribuições individuais para o avanço da tecnologia, esta cresceu e se desenvolveu, na América, especialmente graças à associação entre militares, universidade e firmas.(KIDDER, 1982, p. 13)

O uso, cada vez maior de computadores, inicialmente pelas grandes corporações, mas a partir da década de 80 do século XX, também pelo cidadão comum, representou um grande salto para o registro, armazenamento e recuperação de documentos. Estas máquinas, em função da exigência de cada vez mais espaço para registro de seus *bits* (codificação digital), passaram a utilizar diferentes tecnologias, desde as fitas magnéticas, passando por discos magnéticos, ópticos e diversos outros. Hoje, no início do século XXI, as novidades incorporadas ao conjunto de mídias são os *tocadores de áudio*, *pen-drives* e *iPods*.

O que se observa também é que, a partir dos documentos com conteúdos basicamente textual, encontrados até meados do século XX. Passamos, neste início do século XXI, a encontrar uma grande diversidade de conteúdos, desde imagens fixas, imagens em movimento, registros sonoros, até a combinação de todos estes elementos. Além dos

documentos, cada vez mais importantes, que surgiram especificamente com o uso da informática, como bancos de dados, planilhas eletrônicas e etc.

O DOCUMENTO ELETRÔNICO E DIGITAL

Uma comparação entre os documentos disponíveis e suas características até o final do século XIX com os documentos eletrônicos e digitais atuais, revela as características contrastantes e peculiares dos últimos, elencamos aqui, pelo menos quatro delas:

Legibilidade por Máquina. A necessidade de se utilizar máquinas para que seja possível o acesso ao conteúdo destes documentos nos parece ser a característica mais marcante. Apenas no final do século XIX e durante o século XX existem documentos que não podem ser acessados diretamente pelo homem, sem a ajuda de máquinas.

Independência entre suporte e conteúdo. Tradicionalmente, os suportes físicos da informação registrada em um documento, não podem ser removidos, sem danos ao documento em si. O documento tradicional, além de ser constituído pelo conjunto suporte com informações registradas, forma um conjunto indissociável. O documento eletrônico digital, por sua vez, ainda será formado pela dupla suporte e conteúdo e sempre necessitará de um suporte físico para completar esta dupla, no entanto, percebe-se que o suporte físico pode ser facilmente substituído, sem danos no que cabe ao conteúdo de informações registradas.

Codificação digital. A informação registrada em um documento tradicional, utilizará diferentes linguagens, desde a composição visual e artística até o vernáculo particular de cada grupo social, podemos afirmar que a linguagem utilizada na gravação dos registros de um documento tradicional é variada. Nos documentos eletrônicos e digitais, apesar da visualização de diferentes linguagens quando do acesso ao conteúdo destes documentos, em última análise, a linguagem utilizada na gravação destes documentos será sempre digital, independentemente do vernáculo, tipo de imagem, cor ou características sonoras do documento.

Diversidade de conteúdos. O tipo de conteúdo, até o advento dos documentos eletrônicos e digitais, é diferente para cada tipo de documento, assim temos os documentos com conteúdos textuais, que são diferentes daqueles com conteúdos de imagens fixas (fotografias, por exemplo), documentos fotográficos, ou os documentos sonoros (com conteúdo de áudio) e etc. Por outro lado, os documentos eletrônicos e digitais, formam um grupo único, e são capazes de decodificar, a partir da linguagem digital utilizada para sua gravação, diferentes tipos de conteúdo, como o som, o texto e diversos outros.

PRESERVAÇÃO DE DOCUMENTOS ELETRÔNICOS E DIGITAIS

Percebemos então, através de nosso breve histórico, a diversidade existente de documentos, em diferentes suportes e épocas. Hoje apesar de apenas alguns suportes serem utilizados na produção normal e diária de documentos, como diferentes tipos de papel, CD's ou fitas magnéticas, existem acervos documentais, compostos uma gama muito maior de materiais, este foram sendo formados e acumulados ao longo de séculos e até milênios. Assim, profissionais da área de conservação e restauração, têm de enfrentar problemas com acervos documentais compostos por discos de vinil, discos de acetato, películas fotográficas, diversos tipos de fotografia (além do papel fotográfico comum), papiro e pergaminho (alguns com milênios de idade).

Porém, como Neil Beagrie e Maggie Jones (2001) notaram, os desafios de preservação de documentos digitais estão relacionados às diferenças entre estes e os documentos tradicionais, como a dependência em *hardware* e *software*, a velocidade de evolução tecnológica que implica em cuidados muito mais prematuros, a fragilidade das novas mídias em relação a suportes como o papel ou o microfilme entre outras diferenças.

Quem e porquê preservar este documentos ?. Neil Beagrie (2003, p. 4) observa que as instituições Arquivos e Bibliotecas sempre estiveram a frente do problema da preservação de documentos e são instituições privilegiadas para tomar decisões que afetarão o futuro da sociedade da informação. Várias iniciativas mundiais estão ligadas a este tipo de instituição.

Talvez em função de como as atividades de conservação e preservação de documentos eram desempenhadas com documentos em suportes tradicionais, ou seja, com foco na preservação dos suportes físicos ainda existentes, como o material acetato de um disco, o pergaminho e etc. Inicialmente, muitos consideraram as atividades de preservação de documentos eletrônicos e digitais como sinônimo de manutenção das atividades de armazenamento e manuseio, cuidados com foco no suporte físico destes documentos. Talvez a esta tendência natural tenha se somado a evidente fragilidade dos suportes físicos utilizados com documentos eletrônicos e digitais. Restaurar documentos eletrônicos e digitais, teoricamente é possível, porém, trata-se de uma atividade complexa e cara, Seamus Ross e Ann Gow (ROSS; GOW, 1999) prepararam um trabalho sobre o assunto.

Os cuidados com a manutenção adequada dos suportes físicos, sempre será uma questão relevante, já que mesmo o documentos eletrônicos e digitais, incluindo aqueles com conteúdos em rede, como a internet, sempre necessitarão de suportes físicos, em última

análise. No entanto, como vimos anteriormente, o suporte físico destes documentos pode ser descartado, sem afetar as informações do documento. Além do mais, como veremos adiante, vários outros problemas são muito mais prementes que os cuidados com o suporte físico destes documentos.

OS FORMATOS DE ARQUIVO

Neste artigo, nosso foco não está nos problemas relacionados aos suportes físicos de documentos eletrônicos e digitais. Mas sim, na forma e estrutura como as informações estão gravadas nestes documentos. Diversas propostas de atividades relacionadas à preservação destes documentos têm como base esta forma e estrutura, a qual, no jargão técnico constitui os Formatos de Arquivo.

A partir de um projeto implementado por pesquisadores da Universidade de Leeds no Reino Unido e subsidiado pelo *Joint Information Systems Committee* (JISC), foi produzido um relatório a partir de um *survey* sobre Formato de Arquivo e problemas relacionados, que assim o define:

> No nível mais básico, objetos digitais são seqüências de zeros e uns que representam dados codificados. Diferentes **Formatos de Arquivo** especificam como estes códigos representam o conteúdo intelectual criado por um autor do objeto digital. Um exemplo disto é o formato Microsoft Word. Este formato é uma especificação para armazenamento de dados textuais, bem como informações de formatação. Muitos Formatos de Arquivo são incrivelmente complexos, de maneira que os códigos podem ficam ininteligíveis para um observador humano. Para que este objeto digital tenha sentido, um software será necessário para interpretar e exibir (ou renderizar) os dados para o usuário. (UNIVERSITY OF LEEDS, 2003, tradução e grifo nosso)

Podemos agrupar os Formatos de Arquivo em algumas categorias mais comuns, para efeitos de entendimento, são elas os formatos de (com algumas extensões reais do nome do formato):

- Texto (.txt, .rtf, .doc)
- Imagens fixas (.tif, .jpg, .gif, .png, .bmp)
- Imagens em movimento (.mpeg, .avi)
- Som (.mp3, .wav)
- Bancos de Dados (.db, .sql)
- Planilhas eletrônicas (.wr1, .xls)

Para cada Formato de Arquivo produzido por determinado software, existirá uma especificação técnica (embora, como veremos mais adiante, não necessariamente disponível para o público em geral), na verdade haverá também uma especificação para cada versão de um determinado Formato, por exemplo a especificação TIFF 5.0 e a TIFF 6.0, cada uma, com seu detalhamento técnico. Dependendo do Formato de Arquivo, tal especificação técnica pode ser extremamente diferente para cada versão de um mesmo formato.

As especificações de cada Formato de Arquivo são de caráter bastante técnico e estão no escopo de desenvolvedores de software em geral. Tais documentos, explicam, detalhadamente, como as seqüências de bits no arquivo devem ser estruturadas, onde cada tipo de dado deve ser gravado. Para cada formato de arquivo haverá diferenças marcantes entre as especificações.

A figura 1 mostra um trecho da especificação JPEG versão 1.02, extraída do manual disponibilizado pelas instituições responsáveis pela especificação. Assim como esta, diversas outras especificações estão disponíveis para consulta.

Length	(2 bytes)	Total APPO field byte count, including the byte Count value (2 bytes), but excluding the APPO Marker itself
Identifier	(5 bytes)	= X'4A', X'46', X'58', X'58', X'00' This zero terminated string ("JFXX") uniquely Identifies this APPO marker. This string shall have zero parity (bit 7=0)
Extension_code	(1 byte)	= Code which identifies the extension. In this version, the following extensions are defined: = X'10' Thumbnail coded using JPEG = X'11' Thumbnail stored using 1 byte/pixel = X'13' Thumbanil stored using 3 bytes/pixel
Extension_data	(variable)	= The specification of the remainder of the JFIF extension APP0 marker segment varies with the extension. See below for a specification of extension_data for each extension.

Figura 1 – Parte da especificação JPEG, fonte: manual de 1992

Um ponto crucial sobre Formatos de Arquivo, e que está diretamente ligado a problemas com sua preservação, refere-se ao fato de se tratar de um formato proprietário ou não. Em outras palavras, por diversos motivos, principalmente interesses comerciais, os detalhes técnicos de um formato de arquivo podem não estar disponíveis em momento algum ao público em geral. Sobre este problema:

> Um formato é freqüentemente controlado como propriedade intelectual de uma entidade comercial, a qual, tipicamente tem grande interesse em esconder o código base. A competição direciona freqüentes mudanças no formato individual, tabti quanto nas empresas que os controlam; as tecnologias da informação também impõem contínuas transformações. Esta combinação de opacidade e mudança significa que não há segurança de que a tecnologia futura irá suportar os formatos de hoje. De fato, o cenário digital de amanhã será repleto de objetos grandemente difíceis de preservar, acessar e interpretar. (LeFURGY, 2003, tradução nossa)

Para cada Formato de Arquivo produzido por determinado software, existirá uma especificação técnica (embora, como veremos mais adiante, não necessariamente disponível para o público em geral), na verdade haverá também uma especificação para cada versão de um determinado Formato, por exemplo a especificação TIFF 5.0 e a TIFF 6.0, cada uma, com seu detalhamento técnico. Dependendo do Formato de Arquivo, tal especificação técnica pode ser extremamente diferente para cada versão de um mesmo formato.

As especificações de cada Formato de Arquivo são de caráter bastante técnico e estão no escopo de desenvolvedores de software em geral. Tais documentos, explicam, detalhadamente, como as seqüências de bits no arquivo devem ser estruturadas, onde cada tipo de dado deve ser gravado. Para cada formato de arquivo haverá diferenças marcantes entre as especificações.

A figura 1 mostra um trecho da especificação JPEG versão 1.02, extraída do manual disponibilizado pelas instituições responsáveis pela especificação. Assim como esta, diversas outras especificações estão disponíveis para consulta.

Length	(2 bytes)	Total APPO field byte count, including the byte Count value (2 bytes), but excluding the APPO Marker itself
Identifier	(5 bytes)	= X'4A', X'46', X'58', X'58', X'00' This zero terminated string ("JFXX") uniquely Identifies this APPO marker. This string shall have zero parity (bit 7=0)
Extension_code	(1 byte)	= Code which identifies the extension. In this version, the following extensions are defined: = X'10' Thumbnail coded using JPEG = X'11' Thumbnail stored using 1 byte/pixel = X'13' Thumbanil stored using 3 bytes/pixel
Extension_data	(variable)	= The specification of the remainder of the JFIF extension APP0 marker segment varies with the extension. See below for a specification of extension_data for each extension.

Figura 1 – Parte da especificação JPEG, fonte: manual de 1992

Um ponto crucial sobre Formatos de Arquivo, e que está diretamente ligado a problemas com sua preservação, refere-se ao fato de se tratar de um formato proprietário ou não. Em outras palavras, por diversos motivos, principalmente interesses comerciais, os detalhes técnicos de um formato de arquivo podem não estar disponíveis em momento algum ao público em geral. Sobre este problema:

> Um formato é freqüentemente controlado como propriedade intelectual de uma entidade comercial, a qual, tipicamente tem grande interesse em esconder o código base. A competição direciona freqüentes mudanças no formato individual, tabti quanto nas empresas que os controlam; as tecnologias da informação também impõem contínuas transformações. Esta combinação de opacidade e mudança significa que não há segurança de que a tecnologia futura irá suportar os formatos de hoje. De fato, o cenário digital de amanhã será repleto de objetos grandemente difíceis de preservar, acessar e interpretar. (LeFURGY, 2003, tradução nossa)

O que LeFurgy sintetizou tão bem é o principal problema que relaciona as atividades de preservação de documentos eletrônicos e digitais ao formato de arquivo. O termo obsolescência de software é freqüentemente utilizado para se referir a este problema. Como os dados gravados neste tipo de documento, não podem ser lidos diretamente pelo usuário, mas sim através de equipamentos e software, ao longo do tempo, necessitaremos manter os documentos e todo o aparato necessário para sua leitura, como periféricos de leitura, computador, sistema operacional e programas aplicativos originais, no mínimo. Ou, caso se conheça como foi feita a codificação (como são os detalhes técnicos do Formato de Arquivo), no futuro, será possível ler e renderizar novamente o conteúdo originalmente gravado. Note-se que para isto, é fundamental que se conheçam os detalhes do Formato de Arquivo. Sobre isto:

> A ameaça à era da informação digital ultrapassou o perigo das mídias instáveis e obsolescência de hardware. Os problemas mais prementes confrontando os gestores de coleções digitais são o formato de arquivo e a obsolescência de software (LAWRENCE et al, 2000, p. 1, tradução nossa)

Existem outras propostas para tentar manter o futuro acesso às informações de um documento gravado através de um determinado Formato de Arquivo, como a emulação e a migração, que serão vistas mais adiante. De qualquer, forma, estas propostas também dependem do conhecimento sobre Formatos de Arquivo para que possam ser executadas com sucesso, em maior ou menor grau.

Há também questões sobre direitos autorais, que poderão ser relevantes no futuro, caso se tente desenvolver maneiras de acessar um documento gravado em um formato de arquivo que não seja de domínio público.

Diante de todo este quadro, todos os profissionais que se confrontam com acervos de documentos eletrônicos e digitais, têm diante de si o desafio utilizar ou não formatos de arquivo conhecidos (com padrão aberto) ou proprietários. De qualquer forma, um segundo desafio que ocorrerá no futuro, mas que já deve ser planejado, diz respeito a como identificar qual o Formato de Arquivo de um determinado documento. Apesar de hoje, utilizando-se os programas aplicativos que geram tais documentos, isto parecer óbvio, certamente não será óbvio num futuro com outro sistema operacional e outros aplicativos (que não necessariamente e muito provavelmente não serão compatíveis com a tecnologia atual). Este problema fica maior ainda quanto mais se estende tal futuro. Como identificar então um Formato de Arquivo ? E qual a versão deste Formato de Arquivo ?

IDENTIFICAÇÃO DE FORMATOS DE ARQUIVO

Primeiro, vamos enfatizar melhor o problema da versão do Formato de Arquivo, *"[...] arquivos com dados do Word 6.0 não são legíveis pelo Word 5.0, apesar dos arquivos do Word 6.0 e Word 5.0, ambos representarem documentos de processamento de palavras produzidos por duas versões diferentes do mesmo produto"* (OCKERBLOOM, 1998, p. 1, tradução nossa). Considerando este exemplo, possui a mesma importância a identificação do Formato de Arquivo e a versão utilizada, para efeitos de preservação futura do acesso aos documentos que utilizaram determinado formato.

Brown (2005) observa que, além de ser desejável que o processo de identificação de um Formato de Arquivo seja automatizado, propõe como meio para isto a identificação de uma *assinatura*, ou seja, uma seqüência de bits que representam códigos e que estão presentes em cada tipo de Formato de Arquivo. Estas assinaturas, podem ser externas ou internas.

As assinaturas externas *"abrangem todos os indicadores que estão externos à seqüência de bits do objeto digital, como os data forks do Macintosh e as extensões de arquivo do windows."* (BROWN, 2005, p. 7, tradução nossa)

No entanto, ainda o mesmo autor neste mesmo trabalho, enumera algumas desvantagens na identificação de um Formato de Arquivo pelas assinaturas externas, como extensões do nome do arquivo. Como o fato de que uma extensão *não é necessariamente única* para um determinado formato. Elas não permitem, a *identificação da versão do formato* [apesar de existirem exceções no mercado]. Ou o fato de que extensões *podem ser definidas ou alteradas* pelos usuários. Eu acrescentaria também o fato de que a extensão do arquivo, pelo menos em ambiente de Sistema Operacional Windows, só se torna relevante para a identificação quando um aplicativo correspondente está disponível e instalado no Sistema Operacional, caso contrário o usuário será solicitado a indicar um aplicativo correspondente para utilizar aquele formato de Arquivo.

Por outro lado, as assinaturas internas ao arquivo parecem ser um método bem mais promissor, *"Por definição, a especificação de um Formato de Arquivo impõe uma estrutura específica para o conteúdo da seqüência de bits, que é consistente entre todos os objetos daquele formato"* (BROWN, 2005, p. 7, tradução nossa)

A partir, basicamente, do uso de assinaturas digitais e utilizando-se de diversas técnicas diferentes para extrair dados sobre um determinado arquivo e comparar com uma assinatura previamente armazenada, vários instituições ligadas à questão da preservação de objetos digitais, tem disponibilizado software para identificação de Formatos de Arquivo. Ressalte-se que além do próprio software para identificação de um formato específico, é

necessário um banco de dados, previamente armazenado, com a relação das assinaturas internas de Formatos de Arquivo, para cada versão específica.

Duas iniciativas neste sentido são:

O pacote de aplicativos **JHOVE**, (JSTOR/Harvard Object Validation Environment). O qual é capaz, além de executar a identificação de formatos, também sua validação ("o processo de determinar o nível de conformidade de um objeto com a especificação original") e caracterização (quais são as propriedades deste arquivo e formato, como tamanho em bytes, data de atualização e outras);

E o projeto **PRONOM** que disponibiliza o aplicativo DROID, elaborado em código aberto, este aplicativo bastante simples de utilizar, permite após sua instalação atualizações na base de dados de assinaturas de formatos, de maneira a incorporar futuros formatos de arquivo e/ou novas versões.

Existem outras iniciativas com propostas semelhantes, como alimentar um banco de dados com informações sobre formatos de arquivo, possibilitando consultas sobre informações técnicas, as quais, no futuro podem subsidiar ações de preservação.

Os cuidados com o uso de determinados Formatos de Arquivo, não podem se restringir somente à disponibilização de informações técnicas sobre este ou aquele formato específico, na verdade, tais cuidados devem ser tomados ainda na criação dos documentos eletrônicos e digitais em questão. A partir de uma publicação original de Neil Beagrie e Maggie Jones o DPC – Digital Preservation Coalition, mantém um *handbook* que funciona como um guia para a gestão de recursos digitais objetivando manter o acesso aos mesmos ao longo do tempo. Com relação a Formatos de Arquivo este trabalho enumera algumas recomendações para sua escolha e utilização (DPC, 2006, tradução nossa):

- Utilizar formatos de arquivo não proprietários, com código aberto e bem documentados, sempre que possível;
- Utilizar também formatos de arquivo que foram amadurecidos, tenham sido largamente adotados e se tornaram padrão *de facto* no mercado;
- Identificar formatos que possam ser aceitáveis para os propósitos de transferência, armazenamento e distribuição aos usuários (podem ser distintos);
- Minimizar o número de formatos de arquivo a serem gerenciados, dentro dos possível ou desejável;
- Não utilizar criptografia ou compressão para arquivamento de arquivos, se possível.

EMULAÇÃO E MIGRAÇÃO DE FORMATOS DE ARQUIVO

Existem vária propostas (HAAG, 2003, p. 8) para possibilitar a preservação de documentos eletrônicos e digitais. Como a preservação de todo o hardware e software do ambiente original dos objetos digitais. A impressão em papel do conteúdo de objetos digitais (o que é limitado por características como a interação de páginas web, por exemplo). O encapsulamento do objeto digital, basicamente trata-se de manter o objeto digital juntamente com instruções sobre como utilizá-lo. Estas propostas são o resultado de um ambiente ainda indefinido, em termos de preservação de objetos digitais, e apenas o amadurecimento temporal e técnico definirá qual a melhor ou mesmo se poderão ser utilizadas.

Praticamente todos os trabalhos de pesquisa disponíveis, de alguma maneira, mencionam a emulação e migração como alternativas para a preservação de Formatos de Arquivo, ou pelo menos, como alternativas paliativas contra o processo de obsolescência, tanto de software como de hardware. No que segue, faremos uma breve análise destas duas propostas, relacionando-as com a questão dos Formatos de Arquivo.

Emulação – A idéia por trás da emulação é conseguir recriar o ambiente de hardware e software em um computador do futuro, de tal maneira que seja possível acessar um aplicativo e os arquivos gerados pelo mesmo no passado.

> Emulação evita a necessidade de escrever novo software no futuro para exibir formatos obsoletos. Esta é uma vantagem significativa, já que um formato obsoleto tem de ser conhecido detalhadamente para que se escreva tais programas, o que pode requerer pesquisa extensa e possível engenharia reversa, se o formato em questão não estiver bem documentado. (HAAG, 2003, p. 14, tradução nossa)

Esta proposta, que em princípio parece bastante razoável, na verdade traz problemas sérios de implementação. Na prática, não é fácil desenvolver estes *softwares* emuladores; além do mais, existe uma variedade imensa de aplicativos e versões de sistema operacional. Para complicar, alguns formatos de arquivos são proprietários e não se tem acesso à estrutura do formato, impossibilitando assim o desenvolvimento de um bom emulador.

> Porém, um sério entrave é a complexidade para se desenvolver e manter ferramentas de emulação. No futuro, nós teremos que manter várias ferramentas de emulação e não podemos provar que elas sempre funcionarão em plataformas do futuro (OLTMANS, 2005)

Uma outra linha de atuação para a preservação dos documentos em Formatos de Arquivo é a migração de formatos. Basicamente, trata-se de desenvolver meios para *passar* de um formato de arquivo para outro. Assim, sucessivamente, ao longo do tempo, os documentos em arquivos seriam mantidos atualizados tecnologicamente em termos de *software* e *hardware*.

No entanto, também há problemas nesta abordagem; num relatório sobre o tema, registra-se *"No presente estágio, migração como uma estratégia de preservação digital pode ser caracterizada como um processo incerto gerando resultados incertos"* (LAWRENCE et al, 1999, p. 5).

O mesmo relatório citado acima agrupa três grandes categorias de riscos associados com uma estratégia de migração de arquivos:

O primeiro deles, *Riscos associados com a coleção em geral*, refere-se aos riscos relacionados *"à presença ou ausência de apoio institucional, orçamento, sistemas de hardware e software, e uma equipe para gerenciar o arquivo"*.

O segundo, *Riscos associados com os dados do formato de arquivo*, inclui *"os elementos da estrutura interna do formato de arquivo sujeitos a modificação"*.

E, por último, *Riscos associados com o processo de conversão do formato de arquivo*, ou seja, *"o software de conversão pode ou não produzir os resultados pretendidos; erros de conversão podem ser fortes ou sutis"*.

CONCLUSÃO

O problema da preservação de documentos eletrônicos, digitais ou objetos digitais tem recebido, cada vez mais atenção da comunidade internacional, no Brasil as iniciativas ainda são tímidas comparando-se com o quadro lá fora. As principais instituições à frente de projetos ligados a este tema são Arquivos Nacionais e Bibliotecas ou rede de Bibliotecas, vários consórcios e associações foram criados.

Parece fato que este problema não pode mais passar sem a devida atenção. As organizações, tanto públicas quanto privadas, estão se deparando com objetos digitais não mais legíveis, em função da obsolescência de software ou hardware, por exemplo. Ou estão se deparando com a eminência deste problema.

Por se tratar de um tema relativamente novo, internacionalmente as primeiras iniciativas surgem em meados da década de 90, existem muitas propostas de possíveis soluções ainda em debate. Apenas o processo normal de discussão e pesquisas, inclusive

práticas, proporcionará um quadro mais estável e seguro que sirva de guia para os profissionais envolvidos com tais problemas.

Neste trabalho, procuramos apresentar uma visão geral sobre um tema específico dentro da miríade de temas que têm sido discutidos nacional e internacionalmente, os Formatos de Arquivo. Acreditamos que conhecer melhor este aspecto do grande tema Preservação de Documentos no Universo Digital será fundamental para as políticas de preservação institucional. Em contraste com as atividades de preservação de documentos em suportes tradicionais, baseados em papel, principalmente, os documentos digitais necessitam de medidas preventivas antes mesmo de sua criação, como a escolha adequada de um dos inúmeros formatos de arquivo disponíveis no mercado. Além disto, conhecer quais são os formatos de arquivo em uso institucionalmente, também é fundamental para definir como e os custos envolvidos (inclusive financeiros) para sua eventual manutenção a longo prazo e até mesmo a curto prazo, já que em algumas décadas é comum que formatos e softwares associados já apresentem problemas de manutenção no acesso.

Neste artigo, definimos o que é um Formato de Arquivo e os tipos mais comuns. Estes tipos podem ser associados aos diversos documentos em suportes tradicionais utilizados ao longo da história, que inicialmente abordamos em um breve histórico.

Em seguida, abordamos a necessidade, entraves e como pode ser feita a identificação de um Formato de Arquivo e a versão utilizada. Em se tratando da preservação do acesso ao longo do tema, identificar para determinar o quanto se conhece (se é que se conhece) um formato de arquivo, é uma questão fundamental.

Disto decorre várias recomendações para escolha de um determinado Formato de Arquivo, expusemos algumas. As propostas de emulação e migração, entre outras menos comuns, foram apresentadas com o intuito de contrastar e ligá-las à questão dos formatos.

Finalmente, não poderia ser nossa intenção aqui esgotar o assunto. Dentre outros aspectos que não pudemos desenvolver nesta oportunidade e que parecem ser de fundamental importância para a preservação de objetos digitais estão o uso e definição de Metadados, estes descritores de um documento eletrônico, além de serem utilizados para a descrição do conteúdo do documento, também podem descrever aspectos técnicos que auxiliam na continuidade do acesso ao longo do tempo, como informações sobre os aplicativos utilizados para renderizar um arquivo ou o percurso institucional e histórico do documento. Muito próximo da questão do uso de Metadados está o uso de Repositórios para armazenar e, através de diversas medidas e técnicas, manter o acesso a objetos digitais inseridos em seu conteúdo ao longo do tempo. Ambos são assuntos que merecem aprofundamento.

BIBLIOGRAFIA

ARELLANO, Miguel A. Preservação de documentos digitais. **Ciência da Informação**, Brasília, v.33, n.2, p. 15-27, mai/ago 2004.

BEAGRI, Neil. **Na overview of developments in Austrália, France, the Netherlands, and the United Kingdom and of Related International Activity**. Washington: CLIR e Library of Congress, 2003. Disponível em <http://www.clir.org/pubs/reports>. Acesso em 10 de abril de 2006.

BECK, Ingrid. **Building preservation knowledge in Brazil**. Washington: CLIR, 1999. Disponível em <http://www.clir.org/pubs/reports>. Acesso em 10 de abril de 2006.

BROWN, Adrian. **Automatic format identification using pronom and droid**. National Archives: UK, 2005. Disponível <http://www.pro.gov.uk/about/preservation>. Acesso em 10 de abril de 2006.

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

Digital Preservaton Coalition – DPC. The handbook. Disponível em http://www.dpconline.org/graphics/handbook/. Acesso em 05 de abril de 2006.

DOCTORS, Márcio. **A cultura do papel**. Rio de Janeiro: Casa da Palavra, 1999.

HAAG, Den. **Emulation: Context and currente status**. Netherlands: Testbed, 2003. Disponível em http://www.digitaleduurzaamheid.nl. Acesso em 08/04/2006.

_____. **Migration context and current status**. Netherlands: Testbed, 2001. Disponível em http://www.digitaleduurzaamheid.nl. Acesso em 08/04/2006.

HUNTER, Dard. **Papermaking: the history and technique of na ancient craft**. NY: Dover, 1978.

KIDDER, Tracy. **A alma da nova máquina**. São Paulo: Melhoramentos, 1981.

LAWRENCE, Gregory W. et al. **Risk management of digital information: a file format investigation**. Washington: Council on Library and Information Resources, 1999. Disponível em <http://www.clir.org/pubs/reports>. Acesso em 10 de abril de 2006.

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

MUÑOZ-VIÑAS, Salvador. **Contemporary theory of conservation**. Oxford: Butterworth-Heinemann, 2005.

OLTMANS, Erick, KOL, Nanda. A comparison between migration and emulation in terms of costs. **RLG DigiNews**, NY, v. 9, n. 2. Disponível em : http://www.rlg.org/. Acesso em: 20 fevereiro 2006.

SANT'ANNA, Marcelo L. Os desafios da preservação de documentos públicos digitais. **Revista IP**, ano 3, n. 2, dez. 2001. Disponível em <http://www.ip.pbh.gov.br/revista0302/res_ip0302santanna.html>.Acesso em 13 de abril de 2006.

ROSS, Seamus; GOW, Ann. **Digital archaeology: rescuind neglected and damaged data resources**. Glasgow: HATII, 1999.

SMIT, Johanna; GONÇALVES, Cássia Camargo. Como organizar arquivos fotográficos: **projeto como fazer**. São Paulo: AASP, 2005. Apostila do curso

FISCHER, Steven Roger. **A history of writing**. London: Reaktion Books, 2003.

THOMAZ, Kátia; SOARES, A. José. A preservação digital e o modelo de referência open archival information system (OAIS). **Datagramazero**, Rio de Janeiro, v. 5, n. 1 fev. 2004.