Preservação de Vídeo para Milênios

Linda Tadic

Na região de Dunhuang, ao longo da Rota da Seda no deserto de Gobi, na China, há mais de 700 cavernas esculpidas em montanhas por monges budistas. Essas cavernas, criadas entre os séculos IV e XIV a.C. como atos de devoção, são cobertas por impressionantes pinturas murais retratando sutras budistas, detalhes da vida cotidiana como agricultura, danças e cerimônias, além de iconografias de muitas das religiões do mundo: Paganismo chinês, Budismo, Hinduísmo e Cristianismo. Algumas paredes das cavernas contêm até mesmo grafites de soldados russos aprisionados nas cavernas nos anos 1920.

As cavernas são gerenciadas e conservadas pela Academia Dunhuang, que desde o ano 2000 tem gerado fotografias digitais de alta resolução e documentação em vídeo dos murais das cavernas. A Academia se encontra em uma luta sem fim para conservar as cavernas, que são Patrimônio da Humanidade da Unesco. Várias cavernas desmoronaram ao longo dos séculos, sucumbindo à areia, ao vento e às intempéries. A Academia sabe que em mil anos as fotografias digitais podem possivelmente ser o único resquício de algumas cavernas, de modo que estão desenvolvendo um repositório digital para as imagens, vídeos e dados gerados. Os*files**(1)* digitais têm que ser preservados tão cuidadosamente quanto as próprias cavernas. Eu tive o privilégio de trabalhar com a Academia no desenvolvimento de requisitos para o repositório deles, e fiquei impressionada como a experiência deles se assemelha à dos arquivos que mantêm coleções de fitas de vídeo. Embora o vídeo não existisse há 1.500 anos, as dificuldades em conservar a documentação das cavernas e o conteúdo capturado em vídeo são as mesmas.

A necessidade da digitalização

Quando os estudantes de preservação de imagens em movimento de hoje alcançarem a sua marca de 25 anos de carreira (ou mesmo antes), fitas de

vídeo terão deixado de existir. Não só a fabricação de fitas de vídeo terá cessado, ao mesmo tempo em que o mundo irá completar sua transição para a produção baseada em *files*, mas também a maior parte das fitas de vídeo existentes mantidas em arquivos não terá mais sinais recuperáveis a serem transferidos. Assim como uma caverna Dunhunang que pode vir a ser consumida pelas intempéries, o conteúdo desses vídeos que tiverem a sorte de serem digitalizados irá sobreviver como *file* digital; em essência, o conteúdo do vídeo será "renascido" como um *file* digital, e é o *file* digital que deverá ser preservado para o futuro. Com o vídeo, os arquivistas tiveram que deslocar o foco da preservação da mídia para a preservação do conteúdo.

Fitas de vídeo nunca foram projetadas para serem mídias de armazenamento de longo prazo. Sua curta expectativa de vida (EV) obrigou os arquivos na era pré-digital a migrarem periodicamente de um formato de vídeo para outro, em um esforço no sentido de prolongar a existência do conteúdo. Um agravante da migração repetitiva foi a obsolescência do formato de vídeo, que qualquer arquivo que tenha lidado com vídeo pôde vivenciar. Há vinte anos, transferir para Umatic era o procedimento operacional padrão. O Umatic, atualmente um formato obsoleto, foi substituído pelo BetaSP como o formato analógico de destino padrão. O BetaSP está obsoleto atualmente. Os arquivos começaram a transferir para Betacam Digital, mas apesar de a mídia ter um sinal digital, o conteúdo continua em fita de vídeo, com suas questões de deterioração inerentes; o *file* deve ser extraído da fita e tratado como um *file*digital. Além disso, os arquivos estão começando a considerar o DigiBeta como o próximo formato na fila para se tornar obsoleto. Seu irmão de alta definição, o HDCAM, é frequentemente utilizado por redes de radiodifusão e estúdios para conteúdo HD, mas, do mesmo modo, esse é um formato baseado em fita. A escassez das fitas HDCAM causada pelo terremoto de março de 2011 e pelo tsunami no Japão serviu como um alerta aos criadores de conteúdo para pararem de postergar o inevitável e fazerem a mudança para a produção*medialess* ("sem-mídia").

Formato de destino de preservação digital?

Com a obsolescência do formato de vídeo, com as questões de sua curta expectativa de vida, bem como com o fim iminente do vídeo enquanto uma mídia viável, criadores de conteúdo e gestores de todos os tipos estão reconhecendo que devem adaptar sua produção e seu fluxo de trabalho arquivístico do analógico para o digital. O conteúdo está sendo gerado digitalmente ("nativo digital"), e vídeos analógicos estão sendo digitalizados de acordo com a disponibilidade dos recursos.

O processo de transferência de formatos de vídeos obsoletos para formatos de vídeo que se tornarão obsoletos em breve está sob risco de ser replicado no domínio digital. Fazer a transferência para codecs digitais ou formatos que podem possivelmente se tornar obsoletos no futuro não é um bom modelo a ser seguido.

Um dos assuntos mais controversos no nosso campo é sobre qual deve ser o formato de destino padrão de preservação digital quando da transferência de fitas analógicas para *files*digitais. Alguns acreditam que não deve haver um único padrão apropriado para todos os formatos de vídeo de origem e para todos os arquivos; ao invés disso, devemos focar na utilização de um formato aberto para a fácil migração futura e considerar a infraestrutura da instituição, bem como sua capacidade de preservar os *files* digitais.

Ao selecionar um formato de destino, deve-se levar em conta os fatores de sustentabilidade de formato, como aqueles definidos pela Library of Congress***(2)***. A regra básica é que o formato de preservação deva ser:

- um padrão aberto (não-proprietário; isso inclui *files* recipientes também);
- bem suportado (suporte consistente de hardware e software);
- bem documentado (necessário para que a validação e outras ferramentas possam ser criadas para a verificação do *file*).

De preferência, o *file* deve ser o mais sem compressão que o arquivo possa sustentar. Utilizar um formato aberto com o mínimo de compressão possível vai ajudar um arquivo a migrar os*files* adiante no futuro. Pouca ou nenhuma compressão é também algo mais tolerante à perda de bits, enquanto que a perda de bits em um formato muito comprimido pode resultar em perda de informação ou até mesmo em um *file* corrompido e que não pode ser reproduzido, dependendo de onde ocorreu a perda no *file*.

O requisito de **padrão aberto** significa que existem alguns formatos de *file* ou codecs a serem escolhidos para as conversões de analógico para digital. O mais comum é o sem compressão (YUV 10-bit), seguido pelo JPEG2000. Embora o DV25 tenha compressão 5:1, este e o DVCPRO50 (compressão 3:1) são algumas vezes utilizados para a transferência de fitas VHS ou por arquivos com coleções de vídeo tão grandes que sua infraestrutura não pode suportar o "encodamento" de tudo sem compressão.

Fatores de sustentabilidade para recipientes como o MXF (um formato aberto, a não ser que informação proprietária seja adicionada em seu cabeçalho) e para os dois recipientes proprietários Quicktime e AVI devem ser levados em conta tanto quanto o codec e o formato. A Iniciativa de Diretrizes de Digitalização das Agências Federais (FADGI) está esboçando uma Especificação de Aplicação de MXF Aberto para Arquivamento e Preservação (AS-AP)***(3)***, e o resultado do trabalho do grupo está sendo aguardado com grande expectativa pela comunidade arquivística.

Armazenamento de longo prazo

Ao escolher um formato digital de preservação, a instituição deve levar em conta a sua infraestrutura. A instituição tem como sustentar o armazenamento digital, o tempo gasto com pessoal para migrar e conferir os *files* a cada determinado número de anos, além de ter que atualizar o hardware mais ou menos a cada cinco anos? Nós ouvimos que "armazenamento é barato", mas o armazenamento é apenas uma parte dos custos contínuos com preservação digital. Antes de escolher um formato de preservação de destino, um arquivo

deve estimar o volume de armazenamento necessário para pelo menos cinco anos de expansão, bem como os custos com recursos humanos e infraestrutura. A infraestrutura pode incluir hardware, software, eletricidade, ar condicionado, espaço físico e geradores de segurança.

Não existe mídia do tipo "armazenar e ignorar" para *files* digitais. A preservação digital requer ações dirigidas constantes para migrar adiante tanto a mídia de armazenamento quanto o próprio formato do *file*, os quais podem se tornar obsoletos. Esse trabalho deve ser feito independentemente da mídia de armazenamento utilizada: unidades de discos rígidos externos (HDD), servidores RAID, uma Rede de Área de Armazenamento (SAN) ou uma fita digital como a LTO. Instituições de grande porte frequentemente utilizam um conjunto de estratégias de armazenamento, por exemplo, usando uma SAN juntamente com backup em LTO automatizado. Instituições menores tendem a utilizar HDDs ou servidores RAID independentes, mas está se tornando mais comum a elas também fazerem cópias em LTO para backup utilizando unidades de LTO de slot único. Nenhuma dessas soluções é perfeita, de modo que uma instituição deve pesquisar e entender os prós e os contras, bem como o trabalho envolvido em cada solução de armazenamento.

Uma SAN com backup automático em LTO pode envolver altos custos iniciais com hardware e infraestrutura (eletricidade e AC) e custos menores com pessoal para manutenção permanente. Uma operação de menor porte utilizando HDDs ou um RAID ligado a uma unidade de LTO de slot único terá custos iniciais menores de hardware e infraestrutura, mas envolverá mais trabalho uma vez que o processo de backup pode não ser facilmente automatizado. Considerações adicionais quanto ao uso de LTO como uma mídia de backup envolvem o fato de que a fita é fabricada para ter duas gerações de compatibilidade retroativa de leitura, e uma geração de capacidade retroativa de gravação. Isso significa que fitas LTO3 podem ser lidas nos aparelhos LTO5 atuais (mas não gravadas), mas fitas LTO2 não podem ser lidas ou gravadas neles; o LTO2 é um formato de fita digital obsoleto. Portanto, um arquivo que utiliza LTO está impondo a si próprio a atualização do hardware a cada duas gerações (aproximadamente a cada cinco

anos). Há também a questão do software de backup, que grava e cataloga os *files* para fita, para que estes possam ser recuperados. Se um arquivo envia um conteúdo para uma empresa sem especificar qual software de backup ele utiliza, é possível que a empresa possa retornar a fita LTO com o conteúdo transferido utilizando um software de backup/catalogação que não seja possível de ser lido pela instituição. O LTFS (Linear Tape File System) é um módulo agora disponível em alguns aparelhos LTO5 que faz com que a fita LTO se comporte como um *file* simples armazenado em um disco rígido. Essa é uma evolução promissora no armazenamento de *files* removíveis.

Unidades de discos rígidos externos (HDDs) devem também ser atualizados a cada 3-5 anos. Os HDDs são uma mídia de armazenamento de baixo custo para arquivos menores, mas são notórios pelas falhas entre 3-5 anos de uso. Assim como os *files* em uma fita LTO devem ser migrados a cada duas gerações, os *files* em HDDs também devem ser migrados para novos dispositivos a cada 3-5 anos***(4)***.

Proteção de conteúdo

Independente do armazenamento de longo prazo implementado, uma estratégia ou plano de preservação digital deve ser desenhada para a proteção de conteúdo. No seu modo mais básico, o plano de preservação deve incluir a captura de dados sobre a criação do *file* bem no início do seu ciclo de vida, verificações de imutabilidade do *file* (*checksum*), redundância e dispersão geográfica, e um planejamento de migração de armazenamento (com a migração de formatos de *file* assim que o formato estiver sob risco ou obsoleto).

A informação sobre a geração do *file* é chamada de metadados técnicos; estes podem ser utilizados para preservar o *file* no futuro. Tudo sobre a geração do *file* deve ser capturado: o hardware e o software utilizados no caso de uma transferência do analógico para o digital, ou a câmera/dispositivo utilizado no caso de ser um *file* nativo digital; as características técnicas do*file* (codec, formato, versão, tamanho etc.) e o ambiente no qual o *file* pode ser

transmitido/reproduzido (ex.: qual versão do navegador, qual software e qual versão etc.).

A verificação de imutabilidade do *file* é realizada através de um *checksum*. Um *checksum* é gerado na criação de um *file*, e é executado para a verificação de perda de bits toda vez que um *file* é transmitido de um dispositivo de armazenamento para outro. O *checksum* gera uma sequência alfanumérica que é única àquele *file*; caso o *checksum* não corresponda após a transmissão, então o *file* foi corrompido de algum modo. Os algoritmos de *checksum* mais comuns são MD5, SHA-1 e SHA-2. O MD5 é o *checksum* mais comum, mas é atualmente considerado o menos seguro, de modo que muitas instituições estão mudando para o SHA-1 ou o SHA-2.

A ação mais importante na preservação digital é a dispersão geográfica de múltiplas cópias (redundância). Um *file* pode estar corrompido em uma fita de LTO ou em um HDD, mas estar íntegro em outro dispositivo. Dispositivos de armazenamento não são infalíveis: uma fita pode ser dobrada ou deformada, e o drive de um HDD pode ser danificado por partículas finas ou por vibrações. Caso uma instituição tenha recursos, três cópias são recomendadas, mas o mínimo devem ser duas cópias. Uma cópia de redundância não significa que uma cópia está no porão e outra no terceiro andar; mas significa que as cópias estão localizadas bem distantes umas das outras. Se um incêndio destrói seu prédio, seu conteúdo estaria seguro em outras unidades, como, por exemplo, no Colorado ou na Pensilvânia. Os arquivos podem ter acordos de cooperação nos quais um arquivo armazena cópias de backup para o outro e vice-versa. Esse armazenamento deverá ser seguro para que não haja o risco das cópias serem roubadas ou danificadas.

A migração para novas mídias de armazenamento deve ser programada a cada 3-5 anos, dependendo da escolha da mídia de armazenamento pela instituição. Quanto maior o *file*, mais tempo se gasta para recuperar um *file* do local de armazenamento, verificá-lo e copiá-lo para um novo dispositivo de armazenamento. Se formatos proprietários de *file*/codecs forem utilizados ao invés de formatos abertos, o arco de obsolescência desse formato deve ser

observado de modo que o arquivo possa migrar o formato de *file* adiante, de acordo com a necessidade. Uma base de dados de um arquivo deve identificar o codec e o recipiente de cada *file* para que relatórios sobre formatos sob risco ou obsoletos possam ser facilmente executados. Todas essas ações de migração devem ser monitoradas através de metadados.

Metadados

O trabalho humano pode ser o componente mais caro em qualquer estratégia de preservação digital, e parte do trabalho humano é a geração de metadados. A seção anterior descreveu ações de preservação digital, e metadados devem ser gerados para monitorar cada uma dessas ações. Muitos dos metadados técnicos podem ser extraídos automaticamente de um*file*, mas uma pessoa deve inserir a maior parte das informações descritivas, detalhes sobre como um *file* foi gerado e migrado, além de questões de direitos de propriedade intelectual.

Os componentes fundamentais de um registro de metadados são: descrição (sobre o que é o conteúdo e quem foi o responsável por criá-lo), direitos (quem detém os direitos do trabalho em geral, e quais são os direitos subjacentes de terceiros), dados técnicos e de preservação. Existem padrões de estrutura de dados para metadados descritivos, técnicos e de preservação a partir dos quais uma instituição pode selecionar campos relevantes. Não existe mais uma estrutura de dados abrangente, como o MARC; atualmente, uma instituição seleciona os campos a partir de vários padrões que sejam mais relevantes para a sua coleção e gera um dicionário de dados incorporando os campos. Um dicionário de dados inclui um mapeamento entre os vários padrões para a interoperabilidade com outras coleções e futuras migrações de dados.

Padrões de estrutura de dados úteis para vídeo podem variar desde o básico (Dublin Core), direcionados para a radiodifusão (PBCore e EBU Core), específicos para arquivos de filmes (CEN 15744 e 15907***(5)***), direcionados para distribuição de filmes (SMPTE DMS-1), e obviamente ainda há o

MARC. Padrões de metadados técnicos incluem o RP-210, da SMPTE, além do conjunto de metadados técnicos PBCore e EBUCore. Metadados de preservação de*files* digitais são somente representados pelo PREMIS, que é um padrão que independe do formato. Uma instituição deve tentar criar uma estrutura de dados a mais "granular"***(6)*** possível para facilitar a futura migração de dados. É inevitável que esses dados sejam migrados para novos sistemas várias vezes ao longo de décadas.

Vocabulários controlados (ex.: conjunto de termos para assuntos, nomes e lugares) devem ser utilizados para assegurar que os dados sejam descritos de maneira consistente e que o conteúdo possa ser facilmente recuperado pelos usuários. Vocabulários controlados podem tomar a forma de uma lista de seleção simples, tesauro, taxonomias hierárquicas e anéis de sinônimos***(7)***. Se uma instituição não utiliza um vocabulário controlado padronizado, ela deve criar um interno. O conceito-chave é ser consistente.

No fluxo de trabalho do ciclo de vida dos ativos digitais, a geração de metadados tende a ser realizada por mais de uma pessoa. Uma equipe pode adicionar metadados descritivos básicos no início do ciclo de vida do ativo, outra pode adicionar metadados técnicos, outra pode adicionar informações de direitos autorais, e a biblioteca/arquivo pode adicionar metadados de preservação e talvez vocabulários controlados. Utilizar vários equipes para gerar metadados pode atenuar a carga de trabalho, mas as instituições menores muitas vezes têm apenas um catalogador/bibliotecário dedicado aos metadados.

Tempo

Os arquivos sabem que o relógio está girando contra eles e que muito conteúdo em vídeo será inevitavelmente perdido. Os arquivos tomam decisões difíceis sobre o que são capazes de transferir em termos de armazenamento de *files* e de ações contínuas de preservação. Para salvar ao menos alguma representação do conteúdo, alguns arquivos reconhecem que gerar cópias altamente comprimidas é melhor do que não transferir vídeo algum.

Preservação de vídeo é um processo complexo, e para assegurar que o conteúdo baseado em vídeo irá durar tanto quanto as cavernas de Dunhuang, arquivistas de vídeo precisam se tornar preservacionistas digitais. Estamos vivenciando não apenas uma mudança em como preservar conteúdo em vídeo, mas também uma mudança nas nossas habilidades enquanto preservacionistas e arquivistas.

Glossário

MXF (Material Exchange Format). Este é um padrão SMPTE (SMPTE ST 377-1:2011, Material Exchange Format (MXF) -- File Format Specification). MXF é um formato contêiner ou recipiente que pode conter vários fluxos de bits, como vídeo, áudio e XML. *Files* YUV sem compressão e JPEG2000 são suportados pelo MXF. MXF é o contêiner para os Pacotes de Cinema Digital (DCP) e para *files* gerados pelas câmeras Sony XDCAM e Panasonic P2. Mais aqui.

QuickTime. Formato contêiner ou recipiente de propriedade da Apple, também conhecido pela sua extensão de *file* MOV. Este é o recipiente nativo para o software de edição Final Cut Pro.*Files* YUV sem compressão são suportados pelo Quicktime, mas *files* JPEG2000 não são suportados. Mais aqui.

AVI (Audio Video Interleaved). Formato contêiner ou recipiente de propriedade da Microsoft.*Files* YUV sem compressão são suportados pelo AVI, mas *files* JPEG200 não são suportados.Mais aqui.

LTO (Linear Tape-Open). Uma forma de fita digital que armazena dados. Embora seja uma mídia magnética, a formulação de uma fita de dados difere de uma fita de vídeo. Há alguns poucos competidores no mercado de fita de dados, como DLT e AIT, mas, no primeiro trimestre de 2011, a LTO tinha 87% de *market share*, portanto apenas este produto é mencionado neste artigo.

LTO5 é a mais recente geração de LTO***(8)***, e pode armazenar 3 TB de dados sem compressão ou 1,5 TB com compressão. Veja a evolução do LTO aqui.

RAID (Redundant Array of Independent Disks - Conjunto Redundante de Discos Independentes). Uma configuração de servidor voltada para o modo como um servidor ou uma unidade de disco externa protege os *files*. Níveis são atribuídos por números. RAID0 = não há redundância; o servidor ou HDD é armazenamento puro sem qualquer redundância interna. RAID1 = *files* são "espelhados" (isto é, copiados) de uma unidade interna para uma segunda unidade. RAID2 a RAID6 utilizam distribuição e paridade. Para uma explicação completa, veja o verbete Wikipedia aqui.

Fontes para padrões de metadados citados

- MARC21
- Dublin Core
- PBCore
- EBU Core
- CEN 15907: Film identification - Melhorar a interoperabilidade de metadados - Conjunto de elementos e estruturas; prEN 15907:2009
- CEN 15744: Film identification - Conjunto mínimo de metadados para trabalhos de cinematografia
- Os padrões SMPTE podem ser adquiridos no site da SMPTE.

Notas

- **(1)** Como o termo inglês *file* pode significar, em português, tanto uma instituição arquivística quanto um arquivo eletrônico de dados, optou-se aqui, a fim de evitar ambiguidades, por manter o termo *file* em inglês quando este

se referir a um arquivo eletrônico de dados, e traduzir o mesmo termo para "arquivo" quando designar uma instituição arquivística. (N. do T.)

- **(2)** http://www.digitalpreservation.gov/formats/sustain/sustain.shtml
- **(3)** http://www.digitizationguidelines.gov/guidelines/MXF_app_spec.html
- **(4)** Veja o estudo da Google sobre falhas em unidades de disco. Pinheiro, Eduardo et al. "Failure Trends in a Large Disk Drive Population." Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST'07), February 2007.
- **(5)** "CEN" significa Comité Européen de Normalisation.
- **(6)** No contexto de dados, "granular" significa que existem campos distintos mantendo informações específicas, ao invés de um só campo contendo vários elementos de dados. Por exemplo, utilizar campos separados para "formato", "duração" e "estoque do fabricante", ao invés de aglutinar todos os três elementos em um campo "nota", irá facilitar o mapeamento dos elementos em um nova base de dados.
- **(7)** Para obter orientação sobre a criação de vocabulários controlados, ver: National Information Standards Organization (NISO). Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabulary. ANSI/NISO Z39.19-2005.
- **(8)** No momento em que este texto é traduzido, a LTO encontra-se em sua sexta geração (LTO6). Ver em: http://en.wikipedia.org/wiki/Linear_Tape-Open (N. do T.)