O PDF/A-1 COMO MEIO DE PRESERVAÇÃO DE DOCUMENTOS ELETRÓNICOS A LONGO PRAZO

A criação e a conversão de documentos em formato digital têm colocado, nos últimos anos, diversas questões no seio do mundo da informação, especialmente no que concerne à preservação de longo prazo. Como é que podemos garantir que um documento nado-digital ou digitalizado será corretamente visualizado (isto é, tal como foi originalmente criado) dentro de 20 ou 30 anos? De facto, há que reconhecer que o ritmo de obsolescência das novas tecnologias de informação e comunicação impôs uma rutura relativamente a outros suportes não informáticos, como o papel e o microfilme, em que a inteligibilidade da informação, do ponto de vista intelectual, pode estar apenas dependente da capacidade interpretativa humana e de tecnologia totalmente analógica.

Todavia, também é consabido que os suportes clássicos constituem um obstáculo ao acesso remoto à informação e que o uso e abuso do papel têm dificultado a sã gestão arquivística. Tendo em conta esses senãos, a aceitação e utilização de suportes digitais com fins de arquivo de longa duração têm sido graduais, havendo, ainda assim, consensos positivos em torno de determinados formatos, como é o caso do TIFF (ou TIF), que, com uma estrutura bastante estável, oferece garantias de reprodutibilidade a longo prazo. Podemos acrescentar às vantagens do formato TIFF a característica de ser facilmente transmissível, mas não é, por exemplo, intrinsecamente pesquisável e os ficheiros podem apresentar tamanhos pouco compatíveis com a rápida transferência eletrónica de informação.

Esses aspetos foram, decerto, levados em consideração pela AIIM (*Association for Information and Imagem Management*), pela NEPS (*National Printing Equipment Association)* e pelo *Administrative Office of the U.S. Courts* quando tomaram a iniciativa de submeter à ISO (*Internacional Standard Organization)* um suporte eletrónico de preservação a longo prazo baseado no formato PDF, da *Adobe Systems*.

À partida, o formato PDF oferecia maior versatilidade do que o TIFF, pois, ao invés deste, permite o armazenamento de objetos estruturados (isto é, texto, imagens vetoriais e imagens *bitmap*), a pesquisa no texto e OCR interno, bem como a inserção e classificação automática de metadados (título, data de criação, autor, etc.). Além disso, os ficheiros PDF são mais compactos, exigindo uma fração da memória dos ficheiros TIFF. Apesar de vantajosas, estas características não eram, todavia, suficientes para fazer do PDF comum um formato estandardizado de arquivo eletrónico de longa duração. Além de estar sujeito a uma falta de uniformização (uma vez que, no mercado, muitas são as ferramentas que permitem criar PDFs), o PDF, como a maioria dos utilizadores o conhece, não é necessariamente um formato "self-contained", isto é, não obriga a que, por exemplo, todas as fontes de que o documento precisa para ser visualizado estejam embebidas no próprio ficheiro. O mesmo é dizer que, sob a forma do PDF comum, um documento pode ficar dependente de recursos externos para ser corretamente visualizado, o que, a logo prazo, é um risco, porque não se sabe durante quanto

tempo esses recursos externos vão manter-se disponíveis. Por esta ordem de motivos, o grupo de trabalho liderado por aquelas três entidades, a que se juntaram a *Library of Congress*, a NARA (*National Archives e Records Administration*), a Adobe, a Xerox, entre outros, elegeu um novo formato para a preservação de documentos eletrónicos a longo termo, que veio a ser homologado como norma ISO 19005-1:2005 em setembro de 2005: o PDF/A-1.

O PDF/A-1 baseia-se na Referência PDF 1.4 da *Adobe Systems*, uma plataforma estável que foi implementada no Adobe Acrobat 5, com provas dadas e aceite internacionalmente, mas tal referência teve de ser adaptada a novas exigências, que vão ao encontro de um formato de arquivo consistente e de uso generalizado. O PDF/A-1 caracteriza-se, assim, por ser um formato independente de qualquer plataforma de *software* ou *hardware* que se utilize; por ser autossuficiente, isto é, tudo quanto é necessário para visualizar e imprimir um PDF/A-1 está embebido no próprio ficheiro; por ser autodescritivo, uma vez que pressupõe a existência de metadados; por não ter mecanismos de proteção e de restrição de acesso; por ser disponível ao mercado, visto que qualquer pessoa pode usar a Referência PDF e a Especificação XMP (formato de metadados do PDF) para conceber aplicações informáticas que leiam e criem ficheiros PDF/A-1; finalmente, por reunir condições intrínsecas e extrínsecas que fazem presumir um uso crescente e generalizado.

Além de obrigar a determinadas características (fontes embebidas, independência de cor relativamente à plataforma usada e metadados XMP), o PDF/A-1 não admite encriptação, compressão LZW (por motivos de direitos de propriedade), ficheiros embebidos, referências a conteúdos externos, transparências PDF, multimédia e JavaScript. A assinatura digital é suportada pelo PDF/A-1, desde que as fontes utilizadas estejam embebidas no formato.

Por outro lado, o PDF/A-1 divide-se em dois níveis conformes: o PDF/A-1a e o PDF/A-1b. Enquanto que o PDF/A-1a salvaguarda a estrutura lógica e semântica do documento e a sequência do texto, contemplando, por exemplo, o ajustamento do conteúdo textual em écrans de dimensões diferentes (como é o caso dos PDA), as características do PDF/A-1b asseguram apenas a aparência visual dos documentos eletrónicos, sem garantir a coerência textual dos mesmos. Todavia, as diferenças entre estes dois níveis não têm qualquer significado para os documentos digitalizados, mas apenas para os documentos nado-digitais ou que tenham sido objeto de OCR (*Optical Character Recognition)*. Acrescente-se que nem todas as ferramentas de criação do formato PDF/A-1 podem gerar, indiferentemente, ficheiros PDF/A-1a e PDF/A-1b. Na própria *Adobe Systems*, só a versão 8 do Acrobat é que o faz, quedando-se as versões imediatamente anteriores (7.0.7, 7.0.8 e 7.0.9) pela criação de ficheiros PDF/A-1b.

Um aspeto relevante a ter em conta no formato PDF/A-1 é que o mesmo não constitui um sistema ou estratégia de arquivo, nem tão-pouco exclui outros formatos de arquivo, como é o caso do TIFF. Objetivamente, a Norma Internacional 19005-1 apenas identifica um perfil para documentos eletrónicos que garante a sua inteligibilidade ao longo dos anos, ao arrepio das mudanças tecnológicas, pelo que a utilização do formato PDF/A-1 não dispensa a existência prévia de uma organização de arquivo, da qual, aliás, está dependente e é apenas parte, mas que pode ajudar a tornar mais eficaz. Ao mesmo tempo, oferece razões ao legislador para,

finalmente, começar a considerar a preservação digital como uma alternativa capaz à preservação analógica.

Consagrado como norma ISO em 2005 e encontrando-se em fase de aceitação por instituições como a NARA e pelos Arquivos Nacionais da Suécia, o PDF/A-1 já tem, no entanto, um sucessor à vista, o PDF/A-2, que chegará ao mercado nos princípios de 2009. Com base na Referência PDF 1.6, espera-se que a parte 2 do formato PDF/A já admita a compressão JPEG 2000 e a inclusão de multimédia, além de ser compatível com o standard original.

Que se conheça, a primeira ação de formação em Portugal sobre o PDF/A-1 tomou lugar na Associação Portuguesa de Bibliotecários, Arquivistas e Documentalistas (BAD) em 30 de abril de 2007, sob o nome “O Formato PDF/A como proposta de arquivo de longa duração de documentos digitais”, e irá repetir-se em outubro próximo.