

Arquivo de Documentos Comerciais

A problemática da escolha do formato

por
Gil Loureiro
gil.loureiro@logica.com

Lisboa, 2008

Sumário

Desde a utilização de documentos como meio perene de comunicação, que a sua preservação é uma preocupação. A sua custódia ao longo do tempo tem-se revelado problemática não apenas por razões de conservação, mas também por questões relacionadas com a obtenção e legibilidade.

Na era digital, os documentos electrónicos em tudo se assemelham aos seus correspondentes em papel, sendo obvia a evolução dos suportes, bem como dos mecanismos de manipulação e manuseamento. O requisito de arquivo a longo prazo, depende normalmente da natureza dos documentos mas está normalmente presente. No caso específico dos documentos que suportam relações comerciais, este requisito pode tomar a forma de imposição legal, como por exemplo, ao abrigo do Código do IVA, a grande maioria das facturas electrónicas, têm de ser arquivadas pelo emissor de forma a serem reproduzíveis em qualquer ponto temporal durante um período de 10 anos.

Este artigo pretende alertar para um conjunto de factores que influenciam o arquivo e reprodução dos documentos durante o seu ciclo de vida, devendo ser tidos em conta na selecção do formato do documento.

De forma a sistematizar a análise destes factores tornando-os variáveis de decisão, são apresentados formatos normalizados, bem como formatos utilizados no tecido Empresarial e Administração Pública. Estes são analisados relativamente à forma como endereçam os problemas inerentes a cada um dos factores: dimensão dos documentos, durabilidade do formato, garantia de integridade e funcionalidades adicionais como a pesquisa *fulltext* e a vulnerabilidade a vírus.

Simbologia

B2B Business to Business

B2C Business to Consumer

PC Personal Computer

UMIC Agência para a Sociedade do Conhecimento, Ministério da Ciência,
Tecnologia e Ensino Superior.

WORM Write Once Read Many

Índice

1	Introdução	5
2	Variáveis de decisão sobre o formato a adoptar	9
2.1	Dimensão	9
2.2	Durabilidade do formato	10
2.3	Integridade	13
2.4	Capacidade de Pesquisa	14
2.5	Vulnerabilidade a Vírus	14
3	Conclusão.....	15
4	Referências.....	16

1 Introdução

Para a maioria das organizações, existem aspectos relacionados com a escolha do formato que são óbvios, como a dimensão que lhe está inerente, devido ao forte impacto directo no custo da solução de arquivo. No entanto existe um conjunto de outros factores menos perceptíveis que são tão ou mais importantes, alguns apenas fonte de problemas a longo prazo. Um, é a legibilidade do documento durante todo o seu ciclo de vida no arquivo, i.e., qual a garantia que temos que o documento electrónico arquivado num dado formato é visualizável, ou possível de imprimir num dado ponto no tempo.

Uma experiência que vivi recentemente, que devido ao seu enquadramento temporal pode não ser o mais feliz, mas de alguma forma ilustra o problema, deu-se quanto tive oportunidade de ler as minhas diskettes de 5^{1/4} que utilizava no Amstrad PC. Nestas encontrei ficheiros referentes a relatórios de trabalhos que realizei no 12º ano de escolaridade, com extensão ‘pb1’, onde nem me recordava em que aplicativo os tinha criado, ao pesquisar na Internet verifiquei que se tratavam de publicações do *First Publisher*. Continuando a pesquisa não encontrei um programa que me permita visualiza-los.

É espectável, que um formato como o PDF [1], com a ampla utilização que tem actualmente, não sofra deste problema durante a próxima geração, no entanto já tive oportunidade de presenciar problemas reportados pelo *Adobe Acrobat Reader* ao tentar abrir um ficheiro PDF criado e legível numa versão anterior, felizmente este problema foi corrigido nas versões posteriores do *Reader*.

Para completar o conjunto, temos factores como a:

- Vulnerabilidade do documento a vírus, possibilitando o arquivo de vírus durante um longo período de tempo;
- Capacidade de ser reproduzido com exactidão ao longo do tempo versus a apresentação original, que em alguns formatos é garantido pelo embebimento dos recursos; normalmente chamado como formato duradouro;
- Possibilidade de pesquisa não apenas nos meta-dados, mas em todo o conteúdo do documento;

- Independência do dispositivo de visualização de forma a acompanhar a evolução tecnológica dos dispositivos;
- Garantia de integridade, origem e não repúdio que pode ser obtida com base em assinatura electrónica;
- Independência do sistema operativo de suporte, meios de transmissão e localização dos aplicativos de manuseamento;

Ao analisar as várias áreas das organizações que necessitam de documentos comerciais, surgem mais requisitos que lhe estão intrínsecos. Tipicamente, as organizações com uma base de clientes de massa residenciais e uma menor de empresariais, possuem três cenários diferentes de comunicação com os seus clientes sustentada em documentos, cada com requisitos diferentes:

- Documentos destinados ao clientes residências que são grande massa, denominados de B2C, que tipicamente devem ser legíveis por pessoas;
- Documentos destinados aos clientes empresariais, denominados de B2B, que permitam ser interpretados automaticamente pelos sistemas do cliente;
- Documentos com uma forte intervenção humana na sua criação, normalmente que surgem em processos extraordinários de relacionamento com um dado cliente, como exemplo podemos ter pedidos de compra de muito grande valor, que tem de ser comentado e assinado por vários indivíduos;

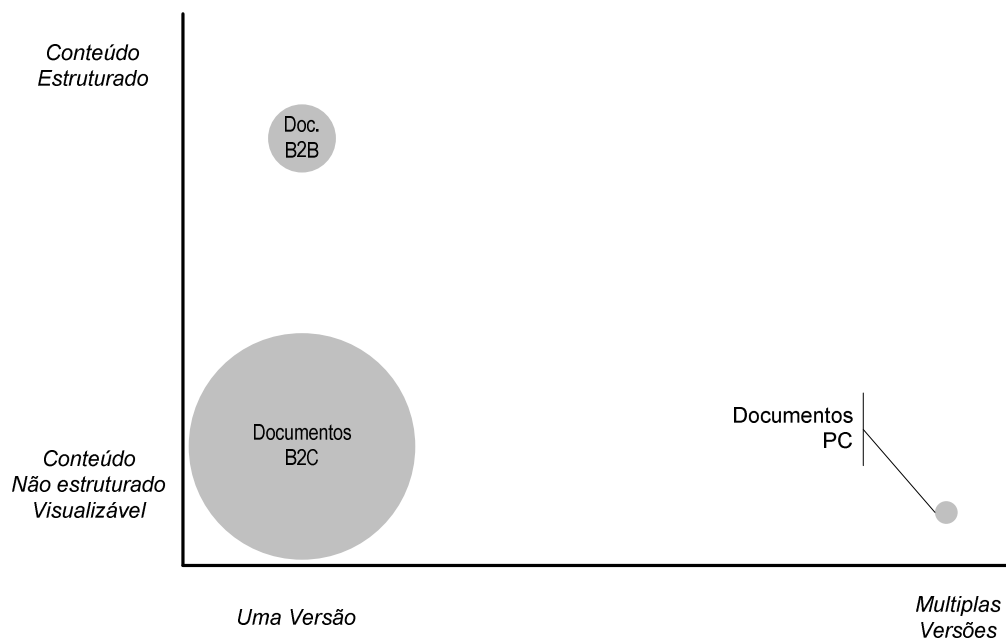


Figura 1 – Tipificação Documentos Comerciais

Na Figura 1 esta representado a volumetria típica de cada tipo de documento quanto à estruturação do seu conteúdo e necessidade de ter mais que uma versão do documento ao longo do seu tempo de vida.

Os documentos B2C são normalmente similares, gerados em massa com base em *templates*, e enviados ao cliente, muita das vezes por obrigação legal a que tenham apenas uma versão. O mesmo se passa com os documentos B2B, que diferem no facto de serem utilizados por sistemas e não pessoas, obrigando-os a serem formatados de forma estruturada (normalmente não legíveis a olho nu) tipicamente sobre XML [3]. Por ultimo temos uma quantidade muito pequena de documentos que são normalmente gerados sobre PC's por vários indivíduos, num processo que todo ele é manual e passa por contribuições de vários indivíduos dando origem a várias versões do documento até que se obtenha o documento final a enviar ao cliente, neste artigo são denominados como documentos PC.

É fácil de perceber que para cada uma destas categorias os requisitos em termos de formato são distintos, tornando difícil ou mesmo impossível que apenas um formato cumpra todos os requisitos.

Neste artigo são detalhados os factores enunciados de forma a simplificar a sua inclusão como variáveis de decisão no processo de escolha do formato de arquivo de documentos comerciais.

No capítulo 2 são apresentados detalhes que ajudam na interpretação de cada variável de decisão, e sempre que oportuno, são referenciado formatos utilizados actualmente e de que forma endereçam o problema.

No capítulo 3 é feita uma análise global com o objectivo de mencionar algumas conclusões genéricas aplicáveis a todas as áreas de negócio.

2 Variáveis de decisão sobre o formato a adoptar

Nas secções contidas neste capítulo são detalhadas cada uma das variáveis de decisão e apresentados casos práticos em que é exposto a forma como determinados formatos de documentos solucionam o problema.

2.1 Dimensão

A problemática do tamanho do ficheiro produzido por um dado formato para representar um documento normalmente apenas se coloca na vertente B2C, devido à sua elevada quantidade. Nesta vertente, garantidamente o PDF é o formato de apresentação *de facto*, devido à larga disseminação dos *viewers* para todas as plataformas, à independência do dispositivo de visualização e pela sua orientação ao papel que facilita a impressão.

Utilizando o mecanismo que permite embeber recursos, como as fontes, imagens, etc., no PDF, é possível obter a mesma representação em qualquer plataforma. Quanto mais recursos se adicionar maior será o tamanho do ficheiro, para minimizar este efeito os *viewers* de PDF obrigatoriamente dispõem de um conjunto de fontes denominado base14, que ao serem utilizadas no texto do documento não necessitam de estar embebidas. Outra funcionalidade é o *font subsetting*, que permite incluir apenas os caracteres da fonte utilizados no texto, minimizando o tamanho das fontes extra base14 embebidas.

Pode também ser utilizada uma estratégia de criar um ficheiro PDF contendo inúmeros documentos, quando se consulta o arquivo extrai-se do ficheiro o documento seleccionado que por sua vez é colocado num novo ficheiro PDF que é retornado. Esta operação é “pesada” computacionalmente devido à estrutura hierarquizada do PDF. Como este tipo de documentos na sua maioria é enviado sobre meio postal, logo impressos, são compostos sobre formatos orientados à impressão como o Postscript [2] e o AFPDS (MO:DCA-P) [4]. Estes formatos são sequenciais, onde os recursos são colocados à cabeça e chamados na sequência de páginas. Com recurso a conversores é possível transformá-los em PDF, então opta-se por arquivá-los e a cada consulta basta

obter o cabeçalho e as páginas correspondentes ao documento, tornando mais leve a operação. Permite também evitar a composição massiva de documentos directamente em PDF.

Obviamente esta estratégia é preferencial à de arquivar cada documento no seu PDF, se a quantidade de consultas ao arquivo permitir ao sistema a construção do PDF a pedido em tempo útil, caso contrário, é essencial o arquivo de cada documento no seu PDF de forma a obter a performance desejada no acesso ao mesmo.

Este factor não é decisivo na escolha do formato dos documentos englobados nas restantes vertentes devido à sua pequena quantidade.

2.2 Durabilidade do formato

É dado o nome de formato duradouro, aos formatos de documentos que no dão a garantia de o conseguirmos ler e/ou visualizar sempre com a mesma apresentação, ao longo de todo o seu tempo de vida. Este requisito é crucial num sistema de arquivo, e a chave para se garantir o cumprimento deste requisito assenta sobre formatos abertos (não proprietários), normalizados por instituições ou associações credíveis, de forma a diminuir o risco de dependência de um único fabricante de software.

Na secção anterior foram mencionados aspectos que elegem o PDF como formato “rei” em B2C. Até meados de 2005, o PDF era um formato de exclusiva propriedade da Adobe Systems, em Outubro do mesmo ano foi padronizado pela ISO um subconjunto da especificação PDF 1.4 denominado por PDF/A (norma ISO 19005-1:2005) [5], que define um formato de documento electrónico para arquivo de longa duração e regras para os seus interpretadores. O PDF/A pretende garantir que um documento é reproduzível sempre da mesma forma ao longo do seu tempo de vida, o que é garantido utilizando uma estratégia de embeber todos os recursos necessários a sua visualização, proibindo a utilização de recursos externos como *hyperlinks*, *script's* em Java, modelos de cor específicos de uma impressora, etc. São definidos dois níveis de conformidade, o PDF/A -1a que garante a reprodução fidedigna da aparência visual do documento e o

PDF/A -1b que requer que a estruturação de documento seja feita de forma a este ser pesquisável.

Muitas organizações que utilizam o PDF para arquivar documentos B2C, não se preocupam em garantir conformidade com o PDF/A, assentando na prática sobre um formato proprietário. Na sua maioria a razão prende-se com o facto das aplicações de geração de documentos em massa não o permitirem, no entanto com alguns subterfúgios e cuidado na sua utilização é possível a obtenção de conformidade, como exemplo, forçar o embebimento de todas as fontes utilizadas, mesmo as base14.

Em B2B, os documentos têm de assentar em formatos estruturados de forma a serem legíveis por sistemas. Desde os anos 80, este requisito era conseguido com base em transacções ponto a ponto UN/EDIFACT (norma ISO 9735) [6]. No início do século, com o *boom* do XML, e das suas vantagens versus o formato hierárquico estruturado do EDIFACT, surgiram inúmeras alternativas de formato de documento estruturado com base em XML, inclusive uma versão XML do EDIFACT (norma ISO/TS 20625) [7].

Em Portugal, no seguimento da resolução Conselho de Ministros n.º 137 de 2005, que visava promover a recepção/emissão de facturas electrónicas pelos serviços e organismos públicos, foi criado um grupo de trabalho orientado pela UMIC, em que participaram os principais fornecedores de soluções de facturação electrónica, com o objectivo de definir quais os formatos de documento deveriam de ser adoptado. O resultado foi a recomendação de dois formatos de documentos: a norma UBL 2.0 da OASIS [8] e a norma GS1XML 2.1 integrado no conjunto de normas eCom da GS1 [9]. Esta iniciativa tinha como alvo o documento factura, no entanto estes dois formatos definem *schemas* de XML para todos os documentos tipicamente utilizados numa transacção comercial, como exemplo, o pedido de compra e o recibo de pagamento.

Os *schemas* de GS1XML foram desenvolvidos com base em modelos de processos do negócio. Esta prática assegura que os modelos da mensagem de negócio sejam de sintaxe neutra e baseada em processos reais do negócio, tornando-o assim um formato flexível.

Este formato obriga a que cada interveniente (o emissor e o receptor) na transacção disponha de um identificador *Global Location Number* e cada produto ou serviço referenciado na transacção disponha de um *Global Trade Item Number* (similar aos

identificadores unívocos de produtores utilizado na produção dos códigos de barras EAN13 existentes em todos os produtos que consumimos no dia-a-dia), existindo um custo associado a cada identificador. Estes identificadores garantem a correcta interpretação dos documentos a nível internacional desde que o sistema seja capaz de interpretar a versão de GS1XML em causa, porque é garantida a não repetição dos identificadores. O UBL é mais rígido e mais complexo em termos de estrutura dos documentos, fazendo com que tenha um acréscimo de esforço na fase de implementação face ao GS1. No entanto, após estar implementado não tem custos no seu tempo de vida como acontece com o GS1, e evita o esforço de gestão de identificadores.

Os documentos criados e capturados em PC's, podem ser suportados em vários formatos, mas deste conjunto existem dois que se destacam por serem formatos abertos e publicados sob normas, o ODF [10] que desde 2006 é uma norma aceite pela ISO (ISO/IEC 26300) e o OpenXML [11] que foi aprovado como uma norma pela ECMA em 2006 (ECMA-376) e submetido a normalização pela ISO em 2007 que continua em curso (ISO/IEC DIS 29500). Ambos os formatos são baseados em XML e permitem trocar não apenas documentos *word* mas também *spreadsheets* e apresentações.

O ODF é um formato construído pela OASIS, baseado no XML criado pela OpenOffice.org, e que conta com uma longa consulta pública a diversos fabricantes de utilitários que permitem manusear o formato. Enquanto o OpenXML é um formato construído unicamente pela Microsoft e que foi desenhado para coexistir com os formatos legados da Microsoft. No entanto a escolha entre os dois não é óbvia, devido à ampla utilização no tecido empresarial do pacote Office da Microsoft por este estar bem integrado com outros produtos da Microsoft muito disseminados, como o sistema operativo Windows e o Outlook. Nas referências [12][13] são apresentados inúmeros aspectos que devem ser tomados em conta na selecção do formato que mais se adequa às nossas necessidades.

2.3 Integridade

Um dos requisitos que todos os sistemas de arquivo devem garantir é a integridade dos documentos ao longo do seu tempo de vida. Tipicamente, existem duas estratégias que podem ser adoptadas. Uma que assenta na guarda em meio físico ópticos não regraváveis (normalmente denominado por WORM), em que depois de ser gravado o documento não é possível altera-lo ou apaga-lo. A outra que é baseada em assinaturas electrónicas, que permitem garantir a autenticidade, integridade e não repúdio dos documentos.

Com o decréscimo sucessivo do custo do espaço em meio magnético, e por este ser muito mais rápido em termos de acesso para escrita e leitura comparativamente com os ópticos, torna-se cada vez mais interessante o segundo cenário. Existem alguns equipamentos baseados em suportes magnéticos, que de alguma forma garantem a característica WORM com recurso a técnicas de hardware, no entanto em algumas áreas de negócio exista alguma controvérsia sobre este tema.

Referenciando o conjunto de formatos já abordados nas secções anteriores, podemos agrupa-los em dois subconjuntos que suportam de forma diferente as assinaturas electrónicas. O primeiro, contém apenas o PDF/A, onde as assinaturas electrónicas são embebidas no documento dentro de um objecto PKCS#7 [15]. A maioria das aplicações que permitem visualizar PDF, no acto de abertura, valida a assinatura contra o conteúdo do documento de forma automatizada e em caso de erro alerta o utilizador. Adicionalmente é possível adicionar uma assinatura visual no documento, que facilita a percepção ao utilizador que está perante um documento assinado digitalmente. A assinatura pode conter também um selo temporal em conformidade com o RFC 3161 [16] garantido a data/hora de geração do documento.

O segundo subconjunto agrupa todos os formatos com base XML; o UBL e o OpenXML suportam assinaturas electrónicas com o formato XML-DSig [17] que facilmente é integrado na sua estrutura XML. O ODF e o GS1XML são omissos quanto ao suporte de assinatura electrónica, na versão 1.2 do ODF é proposto o suporte do XML-DSig (fora da norma ISO), e o GS1XML remete para a fase de transmissão dos documentos a aposição de assinatura electrónica.

2.4 Capacidade de Pesquisa

Quando arquivamos documentos associamos sempre meta-dados que nos vão permitir efectuar as pesquisas. A escolha dos campos a colocar nos meta-dados que cubram o leque de pesquisas feitas pelos utilizadores nem sempre é fácil, e algumas vezes podemos mesmo omitir alguns campos, que tornaram complicadas ou impossíveis algumas pesquisas.

Nos formatos estruturados é obvio que este problema não existe, mas nos formatos de apresentação como o PDF/A, o ODF e o OpenXML este pode ser encontrado. Uma das maneiras de contornar o problema, é permitir a pesquisa livre por todo o conteúdo de texto do documento (denominada por pesquisa *fulltext*), isto apenas é possível quando o texto se encontra em formato vectorial. Todo o conteúdo de texto que é incluído em imagens, por vezes fruto de digitalizações de documentos, fica “invisível” a este tipo de pesquisas. Adicionalmente, os três formatos, permitem o embebedimento de meta-dados em *tags* próprias, o que deve ser utilizada como alternativa de pesquisa de segundo nível evitando possivelmente a alternativa pesada de pesquisa *fulltext*.

2.5 Vulnerabilidade a Vírus

Arquivar por longos períodos de tempo documentos que contenham vírus, faz-nos sempre colocar a questão de se os antivírus serão efectivos quando estes vírus “adormecidos” forem reactivados.

Normalmente, os vírus em documentos recorrem às capacidades de auto execução de *scripts* e *macros*, fazendo com que os formatos normalizados não contemplem a sua inclusão. O PDF/A proíbe expressamente a inclusão de *scripts* e de objectos externos, no entanto no ODF e no OpenXML não estão definidas linguagens de *macro* ou *script* mas permitem a inclusão de ficheiros externos, e estes sim podem conter vírus.

3 Conclusão

Não existem “receitas” que se possam ser seguidas para no final obtermos o(s) formato(s) de documento que devemos adoptar. Cada caso deve ser analisado individualmente, onde devem ser ponderados vários aspectos que podem ter impacto nos documentos durante todo o seu ciclo de vida e não apenas no momento de arquivo. Juntamente com as necessidades dos utilizadores e/ou sistemas e infra-estruturas de suporte, somos levados a escolher um ou mais formatos, de forma a cobrir todos os requisitos. Na vertente B2C dos Documentos Comerciais, para a grande maioria dos casos atrevo-me a mencionar como a escolha acertada o PDF/A.

Tipicamente, os formatos abertos, normalizados por instituições credíveis, já estão preparados para muitos dos problemas mencionados neste artigo, que, a juntar a todos os outros benefícios das normas, torna a sua escolha preferível a formatos não normalizados.

4 Referências

- [1] Adobe Systems Inc., PDF Reference: Version 1.4, Addison Wesley, 2000, ISBN 0201615886
- [2] Adobe Systems Inc., PostScript Reference, Third Edition, Addison-Wesley, ISBN 0-201-37922-8
- [3] Tim Bray, Jean Paoli, C. Michael Sperberg-McQueen, Eve Maler, François Yergeau, John Cowan, Extensible Markup Language (XML) 1.1 (Second Edition), W3C, 2006
- [4] IBM, Advanced Function Presentation: Application Programming Interface Programming Guide and Reference, S544-3872
- [5] ISO 19005-1: Document management — Electronic document file format for long-term preservation — Part 1: Use of PDF 1.4 (PDF/A-1)
- [6] ISO 9735-2:2002 Electronic data interchange for administration, commerce and transport (EDIFACT) - Part 1: Application level syntax rules (Syntax version number: 4, Syntax release number: 1) - Part 2: Syntax rules specific to batch EDI
- [7] ISO/TS 20625:2002 Electronic data interchange for administration, commerce and transport (EDIFACT) - Rules for generation of XML scheme files (XSD) on the basis of EDI(FACT) implementation guidelines
- [8] Jon Bosak, Tim McGrath, OASIS, Universal Business Language v2.0,2006
- [9] GS1 Portugal – CODIPOR ajuda à Implementação da Fatura Electrónica na Administração Pública
- [10] ISO/IEC 26300:2006 Information technology -- Open Document Format for Office Applications (OpenDocument) v1.0
- [11] ECMA-376: Office Open XML File Formats
- [12] Ovitás AS, Research about OpenXML, ODF & PDF, 2007
- [13] Edward Macnaghten, ODF Alliance UK Action Group, Technical Distinctions of ODF and OOXML
- [14] PDF/A Competence Center, TechNote 0006: Digital Signatures in PDF/A-1 2007.
- [15] IETF RFC 2315: PKCS #7: Cryptographic Message Syntax, Version 1.5
- [16] IETF RFC 3161: Internet X.509 Public Key Infrastructure Time-Stamp Protocol (TSP)
- [17] Donald E. Eastlake, Joseph M. Reagle, David Solo, XML-Signature Syntax and Processing, W3C, Recommendation XML DSig, 2002