Interoperabilidade e portabilidade de documentos digitais usando ontologias

Erika Guetti Suca [1], **Flávio Soares Corrêa da Silva**[1]

[1]Instituto de Matemática e Estatística - Universidade São Paulo (IME-USP)
São Paulo – SP – Brasil

{eguetti,fcs}@ime.usp.br

Abstract. *Our purpose is to enable* ***interoperability of documents*** *and achieve* ***portability*** *of* ***digital documents*** *through the reuse of content and format in different plausible combinations. We propose the characterization of digital documents using* ***ontologies*** *as a solution to the problem of lack of interoperability in the implementations of document formats. As proof of concept we consider the portability between* ***ODF (Open Document Format)*** *and* ***OOXML (Office Open XML)*** *document formats.*

Resumo. *Nosso objetivo é possibilitar a* ***interoperabilidade de documentos*** *e atingir a* ***portabilidade*** *simples e confiável de* ***documentos digitais*** *através da reutilização de formatos e conteúdos, em diferentes combinações plausíveis. Propomos a caracterização de documentos digitais usando* ***ontologias*** *como solução ao problema da falta de interoperabilidade nas implementações de formatos de documentos. Como prova de conceito, será considerada a portabilidade entre os formatos de documentos* ***ODF (Open Document Format)*** *e* ***OOXML (Office Open XML)****.*

1. Introdução

As organizações precisam trocar informação através de documentos. Muitas vezes esses documentos são apresentados com formato e conteúdos pré-definidos, que podem ser equivalentes ou quase equivalentes entre si, porém bastantes distintos em diferentes organizações (ou em uma mesma organização em diferentes contextos históricos). Como recurso importante para gerenciar seu conhecimento de forma efetiva e preservar seu capital intelectual, as organizações precisam disponibilizar documentos independentemente do *software* com que foram criados. Propomos a caracterização dos formatos de documentos digitais usando ontologias para favorecer a portabilidade e superar o problema de falta de interoperabilidade de documentos nas organizações.

O trabalho está organizado da seguinte forma: na Seção 2 introduzimos o problema da preservação dos documentos digitais; na Seção 3 apresentamos os conceitos fundamentais da interoperabilidade e portabilidade de documentos; na Seção 4 mostramos os principais conceitos das ontologias; na Seção 5 resumimos alguns trabalhos relacionados. Na Seção 6 explicamos nossa proposta; finalmente, na Seção 7 fazemos as considerações finais.

2. Preservação dos documentos digitais

Um **documento digital** é um documento codificado em formato binário, acessível por meio de um sistema computacional [Gouget et al. 2005]. Nosso trabalho está focado em

documentos digitais criados a partir de aplicações de escritório. As aplicações de escritório são aplicativos voltados para as tarefas de escritório e geralmente estão agrupadas em interfaces de usuário conhecidas como suítes de escritório.

Cada documento digital possui um formato de arquivo. O **formato de arquivo** especifica a estrutura em que os códigos digitais estão organizados [Shepard and MacCarn 2008]. A codificação do formato de arquivo está profundamente relacionada ao programa que o criou. Às vezes, após um período de tempo, se torna extremamente difícil a leitura do documento sem perda significativa da informação. Assim, os documentos podem existir por dezenas de anos, mas a vida útil de uma suíte de escritório não é sempre garantida. Diante disso, as organizações têm que garantir a integridade e perpetuidade dos documentos, mesmo após o *software* que os criou ter desaparecido do mercado [Ngo 2008, Taurion 2009].

Na preservação de documentos é importante provar sua autenticidade. A **autenticidade** é a capacidade de demostrar que um documento digital é aquilo que se propõe ser. É fundamental provar que existe um conjunto de propriedades significativas que foram corretamente preservadas ao longo do tempo. Quanto maior for o número dessas propriedades, maiores serão os requisitos relativos à infraestrutura tecnológica para dar suporte à sua preservação. Torna-se necessária a criação de políticas de preservação que exprimam, para cada classe de objetos digitais, o conjunto de propriedades significativas que serão asseguradas pelo repositório [Ferreira 2006, Rusbridge 2003].

Este trabalho concentra-se na preservação da autenticidade dos documento digitais, mais do que na preservação das características estéticas do documento.

3. Interoperabilidade e portabilidade de documentos

Interoperabilidade[1] é a habilidade de transferir e utilizar informações de maneira uniforme e eficiente entre várias organizações e sistemas de informação. A **interoperabilidade de documentos** é a habilidade das aplicações de documentos de extrair dados de diferentes tipos de documentos e transformá-los em estruturas padronizadas. Esta informação pode ser trocada entre vários sistemas e posteriormente ser processada [Schmidt et al. 2006]. Com a interoperabilidade de documentos baseada principalmente na tecnologia XML, as aplicações podem comunicar-se diretamente com serviços do governo eletrônico.

Por outro lado, a **portabilidade de documentos** é a troca de documentos com todas as informações que eles contêm, principalmente suas configurações de formato [Schmidt et al. 2006]. Na portabilidade de documentos é importante a fidelidade do formato do documento. A **fidelidade de formato** é a capacidade de manter o formato do documento e seu sentido associado, apesar de ser editado em múltiplas aplicações [Ditch 2007].

Ao contrário da portabilidade, a interoperabilidade de documentos está exclusivamente preocupada com a troca da dados corporativos contidos nos documentos e não faz exigências sobre requisitos em termos de aparência, elementos estilísticos, formato ou questões semelhantes [Schmidt et al. 2006].

[1] http://www.governoeletronico.gov.br/acoes-e-projetos/e-ping-padroes-de-interoperabilidade/o-que-e-interoperabilidade

3.1. *Office Open XML* (OOXML) e *OpenDocument Format* (ODF)

OOXML e ODF são os principais padrões abertos de formatos de documentos baseados em XML. No entanto, os dois padrões são incompatíveis e rivais no mercado, provocando a **guerra dos formatos abertos**, gerando discussões técnicas sobre as vantagens e desvantagens de cada um deles.

O **OOXML** foi desenvolvido pela *Microsoft* em 2008 e tornou-se um padrão aberto ISO (ISO/IEC 29500:2008). OOXML foi projetado para representar o corpus preexistente de documentos de processamento de texto, apresentações e planilhas que são codificados pela *Microsoft*. A especificação OOXML contém material normativo e informativo estruturado em aproximadamente 6546 páginas.

O **ODF** foi desenvolvido pela Sun Microsystems em 2002 e seu processo de padronização foi iniciado pela OASIS[2]. O ODF foi projetado para ser uma especificação de formato de documentos independente de fornecedor ou de *software*. Embora a especificação ODF (aproximadamente 700 páginas) seja complexa para os padrões normais, a reutilização de padrões abertos existentes reduz consideravelmente a complexidade da especificação [Eckert et al. 2009, Ngo 2008].

O estudo de [Shah and Kesan 2009] demonstrou que não existem implementações que ofereçam 100% de compatibilidade (avaliação da leitura dos documentos) dentro das implementações de OOXML e ODF.

4. Ontologias

Na ciência da computação, a definição mais citada de ontologia na literatura é de [Gruber 1993]: uma **ontologia** é uma especificação explícita de uma conceitualização. Em 1997, Borst [Borst 1997] ligeiramente modifica a definição de Gruber, afirmando que: uma **ontologia** é uma especificação formal de uma conceitualização compartilhada. A formalização de uma ontologia está definida em cinco componentes: conceitos, relações, funções, axiomas e instâncias [Gruber 1993]. As ontologias, dependendo do seu grau de formalidade, podem ser modeladas baseadas em técnicas de modelagem de inteligência artificial, engenharia de *software* e bancos de dados [Gómez-Pérez et al. 2005].

A ontologia explicita a informação independentemente das estruturas de dados que são usadas para armazenar a informação. Além disso, a conceitualização de uma ontologia pode ser expressa em várias linguagens [Guimarães 2008]. Elas são projetadas para que a informação seja compartilhada entre agentes que garantam compromissos ontológicos. Desse modo, as ontologias viabilizam soluções para problemas como a falta de padronização, falta de interoperabilidade, problemas com a recuperação da informação, falta de reuso, confusões terminológicas, problemas de troca de informações entre agentes de *software*, dentre muitos outros [Uschold and Gruninger 1996].

Este trabalho propõe o uso das ontologias para a modelagem das características da estrutura de formato e do conteúdo corporativo dos documentos digitais.

5. Trabalhos Relacionados

No trabalho de [Eckert et al. 2009] é analisado como os formatos OOXML e ODF especificam as características mais importantes dos documentos e como essas características

[2]http://www.oasis-open.org

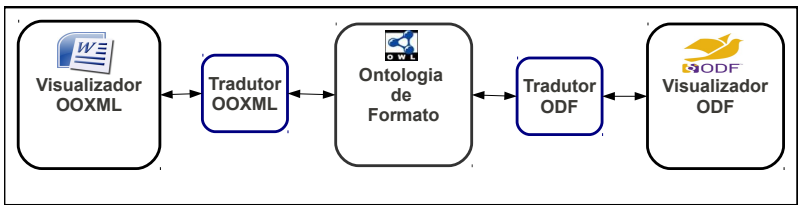

Figura 1. Ontologia como interlíngua.

podem ser traduzidas entre os dois formatos. Esse trabalho concentra-se mais na definição de orientações para a tradução da estrutura da apresentação do documento do que sobre a preservação da estética do documento. Os autores concluíram que a separação da apresentação do documento do seu conteúdo corporativo oferece mais facilidade na manipulação de documentos. Assim, editar os componentes de apresentação e dados do conteúdo corporativo de forma independente confere flexibilidade considerável na criação e edição de documentos.

Outro trabalho é de [Eriksson 2007]. O trabalho dele explica como a combinação de ontologias e documentos cria novas possibilidades para melhorar a gestão do conhecimento nas organizações, isso através dos documentos semânticos. Os documentos semânticos são as integrações dos documentos com as ontologias. O trabalho de [Eriksson 2007] integrou documentos no formato PDF com três ontologias: ontologia de anotação, ontologia do documento e uma ontologia do domínio. Essas ontologias ajudam na explicação do conteúdo do documento e facilitam sua busca. As múltiplas ontologias permitiram ter conceitualizações com diferentes intenções habilitando seu reúso.

6. Caracterização de documentos digitais usando ontologias

Propomos a construção de um modelo que considere as qualidades essenciais de formato de um documento digital. Um documento digital criado com o padrão OOXML ou ODF poderá ser caracterizado nesse modelo.

Nosso modelo será caracterizado a partir de ontologias. Uma **ontologia do formato** especializada em caracterizar a estrutura da apresentação do documento (parágrafos, tabelas, listas, enumerações, etc.), ela apresenta a informação da ontologia de conteúdo. Outra **ontologia de conteúdo** que caracteriza a informação dos dados corporativos contidos no documento (dados do paciente, estudante, vendas, etc.).

A Figura 1 ilustra a interação da ontologia de formato e o documento final. Os documentos digitais serão recriados através de tradutores. Os tradutores são mediadores entre a ontologia de formato e um formato especifico.

A ontologia de formato é criada baseada em um conjunto de propriedades significativas a serem preservadas de uma classe específica de documentos. O objetivo da criação da ontologia de formato é atingir a portabilidade simples do documento, isto é, preservar a estrutura da organização da apresentação do documento, a Figura 2 ilustra os conceitos principais desta ontologia. Por outro lado, o objetivo da criação da ontologia de conteúdo é possibilitar a interoperabilidade de documentos, isto é, habilitar o intercâmbio coerente de informações corporativas específicas sobre um contexto.

A ontologia de formato pode ser modificada independentemente da ontologia de

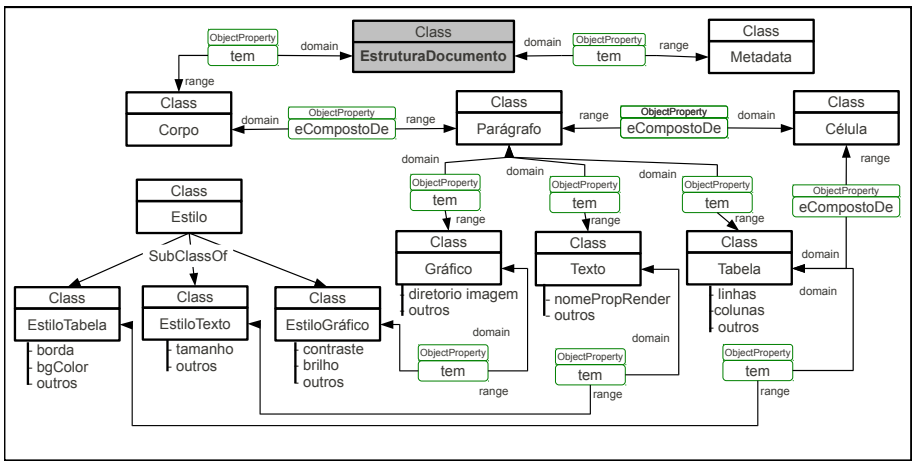

Figura 2. Ontologia de Formato.

conteúdo e vice-versa. Os documentos digitais serão recriados através de tradutores entre a ontologia de formato e um formato específico. Um novo documento pode ser construído com a mesma informação contextual, mas agora num formato apropriado a seus propósitos, bem como um novo documento pode ter a mesma informação contextual, mas com uma apresentação distinta.

6.1. Avaliação dos resultados

No trabalho de [Shah and Kesan 2009] é apresentado um exemplo de avaliação de interoperabilidade e portabilidade de documentos. Eles testaram um conjunto de suítes de escritório que implementam os formatos de documentos ODF e OOXML. Os resultados do estudo estão baseados em pontuações de quão bem as implementações podem ler e escrever documentos. A parte final da pontuação está focado na capacidade de preservação dos metadados dos documentos, isto é, atributos de estilos, números de páginas, tabelas de conteúdos ou cabeçalhos, informações do documento (tempo, ou o número de palavras em documentos), e controle de alterações.

A avaliação dos resultados do nosso trabalho seguirá essa metodologia de avaliação, no entanto com adaptações. Neste caso os documentos serão traduzidos para diferentes formatos, julgando que cada documento pode ser traduzido seguindo vários graus de fidelidade. Para cada documento recriado a partir da ontologia de formato, será quantificada qualquer modificação ao conteúdo original.

7. Considerações finais

A interoperabilidade é um ponto crítico nas questões de governo eletrônico. O presente artigo apresenta a problemática na preservação de documentos digitais e destaca a importância da interoperabilidade e portabilidade nos formatos de documentos digitais. Propomos o uso de ontologias para garantir a integridade e perpetuidade dos documentos digitais, priorizando a preservação das características da estrutura da apresentação e conteúdo corporativo (os elementos importantes para provar sua autenticidade) sobre a preservação das características de estética do documento.

Como trabalho futuro será desenvolvido um caso de uso de interoperabilidade de documentos aplicada ao governo eletrônico, em que as ontologias de formato e conteúdo

tenham papel comprovado na preservação e distribuição eficientes de documentos digitais.

Referências

Borst, W. (1997). *Construction of Engineering Ontologies*. PhD thesis, University of Tweenty, Enschede, NL, Centre for Telematica and Information Technology.

Ditch, W. (2007). *XML-based Office Document Standars*. JISC: Bristol, UK, United Kingdom.

Eckert, K.-P., Ziesing, J., and Ishionwu, U. (2009). *Document Operabilitiy Open Document Format and Office Open XML*. Fraunhofer Verlag, Germany, fokus edition.

Eriksson, H. (2007). The semantic-document approach to combining documents and ontologies. *International Journal of Human-Computer Studies*, 65:624–639.

Ferreira, M. (2006). *Introdução e preservação digital: Conceitos, estratégias e actuais consensos*. PhD thesis, Escola de Engenharia da Universidade do Minho.

Gouget, A. G., Monteiro, B. M., Santos, C. R., da Silva Maçulo, E., de Oliveira, M. I., Miguel, M. L. C., Sobrosa, N. B. S., de Moura Estevão (coord.), S. N., de Mello Lopes, V. L. H., and da Fonseca, V. M. M. (2005). *Dicionario Brasileiro de Terminologia Arquivística*. Arquivo Nacional (Brasil), Rio Janeiro. Edição e Revisão Alba Gisele Gouget and Silvia Ninita de Moura Estevão and Vera Lucia Hess de Mello Lopes and Vitor Manoel Marques da Fonseca.

Gruber, T. R. (1993). A translation approach to portable ontology specification. In *Knowlegde Acquisition*, pages 199–220.

Guimarães, F. J. Z. (2008). Utilização de ontologias no domínio b2c.

Gómez-Pérez, A., Fernández-López, M., and Corcho, O. (2005). *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. Springer-Verlag, 4 edition.

Ngo, T. (2008). Visão geral do office open xml. Technical report, Ecma International.

Rusbridge, A. (2003). Migration on request. *4th Year Project Report Computer Science*.

Schmidt, K.-U., Fox, O., Henckel, L., Holzmann-Kaiser, U., Martin, P., and Tschichholz, M. (2006). *Document Interoperability for Use in eGovernment. Integration of XML-based Document Content in Public Administration Processes*. Fraunhofer Verlag, fokus edition.

Shah, R. and Kesan, J. (2009). Interoperability challenge for open standars: Odf and ooxml as examples. *The proceedings of the 10th International Digital Government Research Conference*.

Shepard, T. and MacCarn, D. (2008). The universal preservation format: A recommended practice for archiving media and electronic records.

Taurion, C. (2009). Adoptando o odf como padrão aberto de documento. Technical report, ODF Alliance Brasil. Volume 1.

Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, methods and applications. *Knowledge Engineering Review*, 11.