Ontologias para descrição de recursos multimídia: uma proposta para o CPDOC-FGV

Daniela L. Silva[1,3], Renato R. Souza [2],

Fabrício M. Mendonça [3], Maurício B. Almeida [3]

[1] Departamento de Biblioteconomia – Universidade Federal do Espírito Santo
Av. Fernando Ferrari, 514 - Goiabeiras – 29.075-910 – Vitória – Brasil

[2] Escola de Matemática Aplicada – Fundação Getúlio Vargas
Praia de Botafogo, 190 – 22.250-900 – Rio de Janeiro – Brasil

[3] Escola de Ciência da Informação – Universidade Federal de Minas Gerais
Av. Antônio Carlos, 6627 – Campus Pampulha – 31.270-901 – Belo Horizonte – Brasil

danielalucas@hotmail.com, renato.souza@fgv.br,
fabriciommendonca@gmail.com , mba@eci.ufmg.br

Abstract. *This paper describes a proposal for building an ontology in the multimedia description domain, in the context of the center for teaching and research in the Social Sciences and Contemporary History (CPDOC) from the FGV. It also presents the results from a state-of-art review study of the multimedia and controlled vocabularies available, and its relation with the Semantic Web Linked Data recommendation.*

Resumo. *O artigo descreve uma proposta para construção de uma ontologia para o domínio da descrição multimídia envolvendo o Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC) da FGV. Apresenta-se também um resultado conciso do estado da arte da temática de vocabulários e metadados multimídia e sua relação com Linked Data.*

1. Introdução

O crescimento exponencial de informações, ocasionado principalmente pelas facilidades introduzidas pelas tecnologias da informação e comunicação, vem impondo desafios no processo de produção, organização e disseminação de informação em Unidades de Informação como Arquivos, Bibliotecas, Museus, Centros de Documentação e Projetos de Memória.

Pesquisas têm sido desenvolvidas progressivamente nos campos das Ciências da Computação e da Informação, visando a estudos sobre a problemática do excesso de informações e sua organização, com o objetivo de melhorar a eficácia dos sistemas de recuperação de informação. Podemos citar, dentre outras, algumas pesquisas nessa perspectiva voltadas à exploração semântica da informação, tais como: a) a Web Semântica e sua proposta emergente de *Linked Data* que intencionam criar

metodologias, tecnologias e padrões de metadados para aumentar o escopo da interoperabilidade e da integração plena de informações heterogêneas entre sistemas de informação [Berners-Lee, Hendler e Lassila 2001] [Berners-Lee 2006]; e b) instrumentos de representação de relacionamentos semânticos e conceituais como ontologias e vocabulários controlados [Gruber 1993], [Guarino 1998], [Silva, Souza e Almeida, 2008] objetivando endereçar problemas relacionados à interoperabilidade de sistemas e bases de dados, além das dificuldades intrínsecas à manipulação da linguagem natural como, por exemplo, as questões de polissemia e sinonímia.

Uma das principais mudanças que reflete a Web é a desterritorialização do documento e a sua desvinculação de uma forma física tradicional como o papel, possibilitando uma integração entre diferentes suportes (texto, imagem, som, vídeo) e a modificação na forma linear de acesso promovida pela inserção das tecnologias hipertexto e hipermídia. Em esfera global, observam-se nos últimos três anos [Schandl et al. 2011] um crescimento significativo de dados semanticamente relacionados e distribuídos na Web – o que se tem denominado na literatura de *Linked Data*. Nesse contexto, padrões de metadados recomendados pelo *World Wide Web Consortium* (W3C) vêm sendo utilizados para descrever e representar recursos multimídia, possibilitando ampliar os pontos de acesso e melhorar a gestão, a organização e a recuperação de acervos digitais. Entretanto, o relacionamento entre multimídia e *Linked Data* ainda é pouco estudado nas comunidades multimídia e ciência da Web [Schandl et al. 2011], abrindo-se oportunidades de pesquisa voltadas a tecnologias eficientes para geração, exposição, descobrimento e consumo de recursos multimídia semanticamente vinculados na Web.

Este artigo objetiva apresentar uma proposta endereçada à construção de uma ontologia de domínio da descrição multimídia para o Centro de Pesquisa e Documentação de História Contemporânea do Brasil (CPDOC). O Centro é dedicado ao estudo e à preservação da memória do país e, atualmente, abriga o mais importante acervo de arquivos pessoais de homens públicos no Brasil (em manuscritos, impressos, fotografias, áudios e vídeos) organizado em sistemas de informações com características próprias. A ontologia de domínio proposta busca a melhoria dos processos de organização da informação do acervo multimídia do CPDOC e a integração de seus sistemas junto a Web de dados.

O presente artigo está estruturado da seguinte forma: na seção 2 são apresentados conceitos, tecnologias e problemas que circundam a temática *vocabulários e metadados multimídia* e seu relacionamento emergente com o paradigma *Linked Data*; na seção 3 é descrita a metodologia para construção de modelos semânticos a um Centro de Pesquisa e Documentação; na seção 4 apresenta-se um resultado parcial de pesquisa sobre vocabulários considerados úteis para o contexto multimídia na Web, e que podem servir para reuso e extensão em um processo de construção de ontologias; e finalmente, a seção 5 é dedicada às considerações finais.

2. Descrição de recursos multimídia e Linked Data

Utilizar metadados é a forma mais comumente empregada para agregar semântica a informações [Gilliland 2000] com o propósito de facilitar a busca de recursos informacionais. No caso de recursos multimídia, os metadados podem ser usados tanto

para descrever atributos técnicos de baixo nível do conteúdo (cores, texturas, timbres de som, descrição de melodia) quanto para descrever características semânticas de alto nível como, por exemplo, classificação de gênero ou representação de informação sobre pessoas retratadas na mídia.

No escopo da Web Semântica [Berners-Lee, Hendler e Lassila 2001], os metadados são agregados através das chamadas linguagens de marcação (do inglês, *markup languages*). Estas linguagens, cujo padrão mais conhecido e utilizado é o XML (*eXtensible Markup Language*), definem *tags* ou marcações que são adicionadas aos dados a fim de indicar alguma informação importante. Ainda que o padrão XML tenha se tornado bastante popular, logo se percebeu que somente esse padrão não é suficiente para permitir a correta interpretação das informações por um sistema informatizado, pois tal sistema não consegue inferir, através das marcações, o que uma informação significa. Tal limitação pode acarretar deficiências nas buscas e na interoperabilidade entre sistemas.

Alternativas estão sendo propostas para este problema pelo W3C no projeto da Web Semântica. Uma dessas alternativas é a adoção do conceito de ontologias para a compatibilização de conceitos encontrados em bancos de dados dos mais diversos tipos na Web. As ontologias apresentam-se como possibilidades de representação de conhecimento em sistemas de informação na medida em que buscam organizar e padronizar conceitos, termos e definições aceitas por uma comunidade particular. Várias linguagens baseadas em XML têm sido propostas para representar ontologias como RDF (*Resource Description Framework*), RDF *Schema* e OWL (*Ontology Web Language*); além da linguagem de consulta para dados modelados em RDF, a SPARQL [Allemang e Hendler 2008].

O enriquecimento semântico sobre dados abertos e vinculados, também conhecido como iniciativa LOD - *linked open data* [Berners-Lee 2006], é uma abordagem recente proposta pelo W3C. A proposta é usar os padrões abertos concebidos pelo W3C em projetos para a Web Semântica a fim de interligar e anotar dados reutilizando vocabulários, ontologias e esquemas de metadados. Nesse sentido, busca-se uma visão integrada de dados e uma maximização da interoperabilidade semântica entre conjuntos de dados (*data sets*) de produtores e consumidores de conteúdo na Web. Os conjuntos de dados Geonames[1] e DBpedia[2] são comumente usados e fazem parte da "nuvem LOD[3]". Entretanto, seus esquemas (além de outros disponíveis na nuvem) não são suficientes para uma atribuição semântica satisfatória aos dados, pois não compreendem um modelo conceitual adequado para representar parte de suas realidades. Além disso, possuem deficiências na qualidade das informações publicadas na nuvem: i) falta de descrição conceitual nos conjuntos de dados; ii) ausência de *links* nos esquemas de dados; e iii) falta de expressividade semântica na representação de dados [Jain et al. 2010].

Provedores de conteúdo multimídia podem enriquecer semanticamente seus esquemas de metadados com especificações estruturadas e bem definidas de

[1] http://www.geonames.org/

[2] http://dbpedia.org/About

[3] Representação gráfica de fontes de dados populares e das ligações entre as mesmas. Cf. http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets/Statistics

conhecimento (por meio de ontologias, por exemplo), viabilizando o consumo e o reuso de informações de alta qualidade e, muitas vezes, multilíngue fornecidas por bases de conhecimento publicamente acessíveis, como o DBpedia, por exemplo. Além disso, podem introduzir *links* para seus descritores de metadados aumentando a visibilidade e a expansão na cobertura de seus conteúdos na Web. Observam-se, assim, mudanças significativas nos modelos de organização e representação do conhecimento no espaço digital no que tange a propostas de melhorar os sistemas de busca e navegação por meio da agregação de abordagens semânticas aos recursos na Web, de forma a obter resultados mais significativos pelos usuários.

3. Metodologia para construção de modelos semânticos para o CPDOC integrados a Web de dados

Uma parte significativa dos conjuntos documentais do CPDOC encontra-se em formato digital e disponível para consulta online. Apesar de poderem ser acessados através do mesmo portal, possuem interfaces e processos de descrição e publicização distintos. São cerca de: 1,2 milhão de documentos manuscritos e impressos (ou 5.1 milhões de páginas); 80 mil fotografias; 6 mil horas de entrevistas em áudio e vídeo; e 8 mil verbetes de natureza biográfica e temática.

O CPDOC conta hoje com um projeto de integração de dados de seus sistemas de informação visando à criação de um portal semântico com interface única para buscas temáticas transversais e integradas. Foi engendrado para promover uma maior integração das bases de dados internas com as externas, como a própria Wikipédia, com benefícios no sentido de aumento da publicização e estruturação de redes sociais de colaboração para contribuições e eventuais correções para o acervo. O projeto prevê a criação de ontologias para descrição de recursos multimídia (áudio, vídeo, imagem, texto) e ontologias no domínio de história contemporânea.

No que tange à ontologia para o domínio da descrição multimídia, foco desta proposta de trabalho, o propósito é conceber modelos conceituais ontologicamente consistentes e bem fundamentados, isto é, dando ênfase à explicitação na semântica dos esquemas de dados internos e externos de interesse do CPDOC. Uma ontologia de domínio bem fundamentada é um modelo de domínio específico que se articula com um domínio de sistema de categorias formal e independente, denominado ontologias de fundamentação [Guizzardi e Wagner 2009]. As categorias ontológicas podem ser úteis no sentindo de esclarecer o significado pretendido dos termos adotados por meio de um conjunto de distinções semânticas, evitando ambiguidade e melhorando, principalmente, a qualidade na representação de dados no contexto *Linked Data*.

A proposta é construir a ontologia de domínio da descrição multimídia orientada por uma ontologia de fundamentação como, por exemplo, a *Unified Foundational Ontology* (UFO) [Guizzardi e Wagner 2009] e a *Descriptive Ontology for Linguistic and Cognitive Engineering* (DOLCE) [Masolo et al., 2003], observando-se, ainda, vocabulários e metadados multimídia disponíveis na Web com vistas a reuso ou a extensão. A ontologia de domínio bem fundamentada para descrição multimídia será útil para a integração semântica entre as bases de dados do CPDOC e estas, por sua vez, ligadas com conjuntos de dados pertencentes à Web de dados. Tal integração será estabelecida por meio de modelos conceituais dos conjuntos de dados envolvidos

ligados à implementação da ontologia de domínio. Acredita-se que a utilização de um nível conceitual é relevante no sentido de abstrair características tecnológicas, além de fornecer uma descrição conceitual para conjuntos de dados e melhorar a compreensão humana e a atribuição semântica às máquinas.

4. Resultados parciais

O objetivo da presente seção é apresentar e descrever de modo sucinto alguns vocabulários (incluindo ontologias) que foram desenvolvidos nos últimos anos pelas comunidades de Web Semântica e *Linked Data*, os quais se mostram relevantes no contexto de marcação semântica para conteúdos multimídia. Tais vocabulários são considerados uma boa prática para reuso ou extensão [Schandl et al. 2011]. O Quadro 1 exibe os vocabulários. Para a exploração da literatura sobre vocabulários e metadados multimídia utilizou-se da técnica de pesquisa bibliográfica e documental em artigos científicos, livros e relatórios técnicos de pesquisa. Para a identificação de documentos relacionados à temática, foram consultadas bases de dados de documentos científicos no portal de periódicos da Capes e na biblioteca digital *Citeseer*. No que diz respeito ao portal de periódicos da Capes, as editoras consultadas foram: i) *Association Computing Machinery*; ii) *Journal Multimedia Tools and Applications;* e iii) *IEEE MultiMedia*.

Quadro 1: Vocabulários relevantes para o contexto multimídia

Vocabulário	Característica
Dublin Core	Fornece propriedades para descrever artefatos criados pelo homem como proveniência, formato, idioma, direitos autorais. Voltado ao domínio de metadados bibliográfico.
Friend of a Friend	Descreve pessoas, organizações e relacionamentos entre eles.
Basic Geo Vocabulary	Define propriedades para a representação de coordenadas geográficas (latitude, longitude e altitude).
Creative Commons	Fornece termos e classes para representar informação legal sobre obras, licenças associadas e permissão de distribuição e uso.
Review Vocabulary	Fornece termos que representam revisões, críticas e comentários para objetos arbitrários.
Multimedia Metadata Ontology	Fornece um *framework* para a integração de aspectos centrais de metadados multimídia.
Core Ontology for Multimedia	Fornece primitivas para explicitar a composição de um objeto mídia e o que nele deve ser representado. É considerada uma ontologia bem fundamentada para descrição multimídia.
Exif Vocabulary	Especifica formatos a serem usados para imagens e sons em câmaras digitais.
Visual Resources Association	Fornece uma organização categórica para a descrição de trabalhos ligados a cultura visual bem como imagens que os documentam.
Categories for the Description of Works of Art	Descreve objetos de arte e imagens, além de incluir discussões e assuntos relacionados à construção de sistemas de informação no domínio da arte.

Segundo [Schandl et al. 2011], existem muitos vocabulários relevantes para dados multimídia, entretanto, ressaltam que uma grande parte ainda não é utilizada no contexto de *Linked Data*.

5. Considerações finais

Este artigo permitiu evidenciar que há uma quantidade considerável de padrões de metadados, vocabulários e ontologias na tentativa de melhor representar recursos

multimídia visando recuperação semântica através de bibliotecas, portais e bases de dados digitais abertos.

Esforços na construção de ontologias podem ser poupados tendo em vista a exploração de vocabulários em comunidades de interesse. Contudo, surgem desafios na identificação e seleção de uma variedade de padrões de metadados, vocabulários e ontologias disponíveis e que precisam ser compatíveis com as entidades reais de um domínio específico. Tais desafios encontram-se i) no alinhamento de vocabulários e ontologias que reflete aspectos de interoperabilidade semântica e sintática para o provimento de compartilhamento entre sistemas e aplicações na web; e ii) na modelagem conceitual adequada para representar consensualmente parte da realidade de um domínio.

Referências

ALLEMANG, D.; HENDLER, J. (2008) *Semantic web for the working ontologist*: *modeling in RDF, RDFS and OWL*, Elsevier, MA, USA.

BERNERS-LEE, T; HENDLER, J.; LASSILA, O. (2001) "The Semantic Web". *Scientific American*, vol. 284, nº. 5, maio, p. 34-43.

BERNERS-LEE, T. (2006) "Linked Data - Design Issues". Available at: < http://www.w3.org/DesignIssues/LinkedData.html >.

GILLILAND, Anne J. (2000) "Introduction to metadata: setting the stage". Available at:<http://www.getty.edu/research/publications/electronic_publications/intrometadata /setting.pdf >.

GRUBER, T. (1993) "What is an Ontology?" Available at: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>.

GUARINO, N. (1998) "Formal ontology in information systems". Available at: <http://citeseer.ist.psu.edu/guarino98formal.html>.

GUIZZARDI, G.; WAGNER, G. (2009) *Using the Unified Foundational Ontology (UFO) as a foundation for general conceptual modeling languages*. Springer-Verlag, Berlin.

JAIN, P.; HITZLER, P.; YEH, P.; VERMA, K.; SHELT, A. (2010) "Linked Data is Merely More Data". *Semantic Technology Conference*. Available at: < http://knoesis.wright.edu/library/publications/linkedai2010_submission_13.pdf >

MASOLO, C.; BORGO, S.; GANGEMI, A.; GUARINO, N.; OLTRAMARI, A. (2003) *Ontology Library*: WonderWeb Deliverable D18. Trento, Italy. Available at: <http://www.loa-cnr.it/Papers/D18.pdf>.

SCHANDL, B.; HASLHOFER, B.; BÜRGER, T.; LANGEGGER, A.; HALB, W. (2011) *Linked Data and multimedia: the state of affairs*. Multimedia Tools and Applications, online first,1-34.

SILVA, D. L. da; SOUZA, R. R.; ALMEIDA, M. B. (2008) "Ontologias e vocabulários controlados: comparação de metodologias para construção". *Ciência da Informação*, v. 37, n.3, p. 60-75.