



XVI Encontro Nacional de Pesquisa em Ciência da Informação (XVI ENANCIB)
ISSN 2177-3688

GT 2 – Organização e Representação do Conhecimento
Pôster

**RECUPERAÇÃO DA INFORMAÇÃO EM ACERVOS DIGITAIS DE
JORNAIS IMPRESSOS: PROPOSTA PARA MODELAGEM DE
ONTOLOGIA NO DOMÍNIO DO FUTEBOL¹**

***INFORMATION RETRIEVAL IN DIGITAL COLLECTIONS OF PRINTED
NEWSPAPERS: PROPOSAL FOR ONTOLOGY MODELING IN THE
FOOTBALL***

Luana Carla de Moura dos Santos, UFSC
luana-moura@hotmail.com

Marisa Bräscher, UFSC
marisa.brascher@gmail.com

Resumo: O objetivo desta pesquisa é aprimorar a recuperação da informação em acervos digitais de jornais por meio da modelagem de um domínio para aplicação de ontologia. O domínio a ser explorado é o futebol. A escolha justifica-se devido à escassez de publicações em outras fontes que retornem informações detalhadas de jogos, jogadores, campeonatos, clubes e da história do futebol em geral. Trata-se de uma pesquisa ainda em desenvolvimento, caracterizada quanto à natureza como aplicada, quanto aos procedimentos técnicos como pesquisa exploratória e descritiva, com abordagem qualitativa. Como resultados espera-se demonstrar como a modelagem de um domínio específico, o futebol, pode minimizar os impactos causados pela ambiguidade e aqueles derivados das características peculiares dos jornais, marcados pelas variações linguísticas tanto diacrônicas quanto sincrônicas.

Palavras-chave: Jornal. Acervo Digital. Ontologias. Futebol.

Abstract: The objective of this research is to improve information retrieval in digital collections of newspapers through modeling a domain ontology for application. The area to be explored is football. The choice is justified because of the scarcity of publications from other sources that return information about games, players, leagues, clubs and the history of football in general. This is a research still in development, characterized by the nature applied as the technical procedures as exploratory and descriptive research with a qualitative approach. The results are expected to

¹ O conteúdo textual deste artigo, os nomes e e-mails foram extraídos dos metadados informados e são de total responsabilidade dos autores do trabalho.

demonstrate how the modeling of a specific domain, the football, can minimize the impacts caused by ambiguity and those derived from the peculiar characteristics of newspapers, marked by linguistic variations both diachronic as synchronous.

Keywords: Journal . Digital Collection. Ontologies. Soccer.

1 INTRODUÇÃO

Assim como o surgimento da prensa de Gutenberg viabilizou a multiplicação dos documentos impressos, a expansão da Internet estimulou as publicações em meio digital. Nessa transição, do impresso para o digital, suscitada pelo avanço significativo das Tecnologias da Informação e Comunicação (TIC), foi evidenciada a necessidade de disponibilizar digitalmente os documentos produzidos em suporte de papel, em décadas e séculos passados, principalmente aqueles provindos de fontes primárias.

No início deste século, jornais brasileiros que mantinham internamente seus bancos de dados, disponibilizaram esses acervos digitais na Internet. O jornal em ambiente digital oferece vantagens quando comparado à versão convencional. O acesso não é limitado, permitindo que vários leitores possam fazer uso simultaneamente de um mesmo documento. A guarda e preservação são facilitadas pelos bancos de dados dos próprios jornais e a pesquisa é simplificada pelo auxílio de recursos que propiciam a busca de termos nos textos.

Em ambientes digitais, de maneira geral, utiliza-se a linguagem natural como linguagem de busca. Por essa razão, explorar e direcionar as ferramentas de busca para o uso linguagem natural é determinante para um sistema de recuperação da informação (RI).

Outra questão que se coloca do ponto de vista da RI em acervos digitais que reúnem documentos de diferentes épocas e culturas é a variação linguística que ocorre em função da forma diacrônica “(dizer já dito em momentos diferentes)” e sincrônica “(dizer atual e simultâneo)” (MOURA, 2009, p. 64), assim como aqueles referentes aos contextos de uso, marcados muitas vezes por aspectos geográficos e socioculturais.

Em conformidade com benefícios apresentados com a aplicação de ontologias em SRI na Internet, surge o interesse de empregá-las no contexto de acervos digitais de jornais e analisar o potencial que podem oferecer para o refinamento dos resultados obtidos na RI desses ambientes.

O termo ontologia surgiu inicialmente na Filosofia, originado do estudo da metafísica. No entanto, a pesquisa sobre ontologias tem aparecido também com frequência na literatura da Ciência da Computação e da Ciência da Informação. Na Ciência da Computação o termo passou a ser usado inicialmente nos anos 1960, em pesquisas relativas à Representação do Conhecimento, subárea da Inteligência Artificial e nos anos 1980 surgiram as primeiras

pesquisas relacionadas à Engenharia de Software. Nos 1990 as publicações dos pesquisadores da Ciência da Informação demonstravam um crescente interesse sobre o assunto. Nessa visão as ontologias são identificadas como um tipo de sistema de organização do conhecimento, com o propósito de organizar e representar o conhecimento (SOERGEL, 1997; VICKERY, 1997; ALMEIDA, OLIVEIRA, COELHO, 2010; ALMEIDA, MENDONÇA, AGANETTE, 2013).

Na proposta da Web Semântica as ontologias surgiram como alternativa para agregar semântica aos recursos da web (BERNERS-LEE, HENDLER e LASSILA, 2001). De acordo com Breitman (2010, p. 7) por meio de “modelos conceituais que capturam e explicitam o vocabulário utilizado nas aplicações semânticas” as ontologias podem contribuir para uma “comunicação livre de ambiguidade”.

Um exemplo que ilustra o papel da ontologia é o uso do termo “banheira”, que de acordo com o Dicionário da Língua Portuguesa - Aurélio², pode tanto ser entendido como “tina para banho” e também como “posição irregular de um jogador quando, no momento do passe de um jogador da mesma equipe, se encontra no meio-de-campo adversário, à frente de todos os jogadores da outra equipe, à exceção de um”. Essa diferenciação é compreendida facilmente por humanos, mas máquinas não conseguem diferenciar os significados conforme o contexto, se não tiverem o auxílio de recursos que permitam identificar esses significados.

No exemplo da frase “Paulo acertou uma bela bicicleta”, se a sentença tiver relação com o domínio do futebol, é possível concluir que a palavra “bicicleta” se refere à jogada em que o atleta chuta a bola com o corpo no ar. De acordo com Studer, Benjamins e Fensel (1998, p. 184) “a ontologia pode ser compreendida como um modelo abstrato de algum fenômeno no mundo, esse fenômeno pode ser construído por meio da identificação de conceitos relevantes, geralmente um conhecimento do domínio”.

No contexto dos acervos digitais de jornais, as mudanças e variações linguísticas ocorridas durante a passagem das décadas também precisam ser avaliadas, como no exemplo das palavras “Directoria” e “Thesoureiro”, que tiveram alterações na sua grafia, ou das palavras “escanteio” e “futebol”, que substituíram o uso dos termos em inglês “corner” e “foot-ball”, anteriormente utilizados no Brasil. Desse modo, é preciso criar relações semânticas entre esses termos, para que as distintas formas de grafia sejam recuperadas, sem que seja preciso realizar inúmeras buscas no sistema.

² DICIONÁRIO DA LÍNGUA PORTUGUESA - AURÉLIO. **Significado de ‘banheira’**. 2015. Disponível em: <<http://www.dicionariodoaurelio.com/banheira>>. Acesso em: 8 jul. 2015.

Assim sendo, à medida que as tecnologias da Web Semântica avançam e oferecem novas possibilidades de RI, se aplicadas no contexto dos acervos digitais, também podem potencializar o acesso às informações.

O domínio explorado nesta pesquisa é o do futebol. A escolha justifica-se devido à importância dos acervos digitais de jornais como fonte de informação sobre esse domínio, pois disponibilizam informações detalhadas de jogos, jogadores, campeonatos, clubes e da história do futebol em geral.

2 PATRIMÔNIO DIGITAL

Para Dodebei (2006), patrimônio pode ser considerado como um atributo simbólico de um bem que se deseja preservar com indício de memória. Desde a antiguidade, a ciência histórica reúne documentos escritos e faz deles testemunhos (LE GOFF, 1990). O jornal, nesse sentido, é um indício de memória, que por meio da escrita, fornece informações sobre a história de uma sociedade.

No contexto digital, o patrimônio, de acordo com Dodebei (2006), pode ser estudado sobre dois enfoques: o primeiro considera o patrimônio digital como um processo de constante produção, seja dos objetos digitalizados, ou dos objetos já que já nasceram digitais (*born-digital heritage*). O segundo refere-se ao patrimônio como produto de uma escolha e, nesse enfoque, um conjunto isolado de objetos é escolhido para representar, de forma simbólica, as ações sociais.

Os acervos digitais de jornais estão em conformidade com o primeiro enfoque, pois disponibilizam em suas bases tanto edições de jornais que foram criados e distribuídos em papel, quanto edições que já nasceram digitalmente. Ambos os formatos se integram e formam uma linha do tempo de informações históricas que possuem data de início, assinalada pelos primeiros exemplares dos jornais, mas não possuem data final, pois a história continua e a reprodução dos jornais segue concomitante. O segundo enfoque também é colocado em evidência e relaciona-se com a salvaguarda da integridade dos acervos. No sentido digital, além da preservação, as formas de representação que garantem o acesso também devem ser contempladas.

No que se refere aos acervos digitais, a virtualização e a digitalização são demandas práticas e cotidianas no processo de transferência da massa documental física para ambientes digitais (DODEBEI, 2006). Contudo, os novos suportes carecem de intervenção para o desenvolvimento de mecanismos que propiciem um acesso facilitado a essas mídias e, nesse

sentido, explorar medidas que visem o uso contínuo de informações, para que elas não se tornem inacessíveis, também é fundamental.

3 JORNAL: DO IMPRESSO AO DIGITAL

No Brasil, o jornal impresso nasceu no início do século XIX (1808) com finalidades políticas, diferentemente dos interesses de fortalecimento da classe mercantil, demonstrados no cenário Europeu, no início do século XVII, e nas Américas espanhola e inglesa, ao final do mesmo século (CALDAS, 2004).

No início do século XX, o jornal ainda era o principal veículo de comunicação da sociedade, fato que culminou o título de anos dourados à mídia (ANJ, 2008)³. Nos anos 1920, surge o rádio e após a televisão, no entanto o folhetim diário não se tornou obsoleto.

Finalmente nos anos 1990, os jornais adentraram no contexto digital e assumem, assim, outras características, mas não perdem a sua identidade social e cultural. Os acervos digitais de jornais são formados por jornais digitalizados e jornais que já nasceram digital, que compõem o chamado “patrimônio nascido digital” (*born-digital heritage*), ou “nascidos digitais”, e possuem as mesmas características da versão impressa, com uma única publicação diária. Juntos (digitalizado e “nascido digital”) constituem os acervos digitais e são os objetos de estudo desta pesquisa.

4 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa de mestrado, em fase de desenvolvimento, caracteriza-se, quanto à natureza como aplicada, haja vista a sua orientação para a resolução de problemas voltados à aplicação prática. Quanto aos procedimentos técnicos caracteriza-se como uma pesquisa exploratória e descritiva, com abordagem qualitativa.

As etapas consideradas para o alcance dos resultados desta pesquisa foram agrupadas em três partes: i) a primeira está concentrada na estruturação do corpus de pesquisa;; ii) a segunda parte relaciona-se com a modelagem do domínio do futebol iii) e a terceira destina-se à construção da ontologia.

Na elaboração da ontologia do domínio do futebol, a ser desenvolvida neste trabalho, será utilizada a metodologia proposta por Mendonça (2015) – OntoForInfoScience. As etapas consideradas são: Especificação da ontologia; Aquisição e extração do conhecimento; Conceitualização; Fundamentação da ontologia; Formalização da ontologia e Avaliação do

³ ASSOCIAÇÃO NACIONAL DOS JORNAIS. 2015. **Jornais:** Breve História. Disponível em: <<http://www.anj.org.br/jornais-breve-historia-2/>>. Acesso em: 8 jul. 2015

conteúdo ontológico. As etapas que envolvem a documentação e disponibilização da ontologia, também contempladas na OntoForInfoScience, não serão consideradas nesta pesquisa, devido ao curto período disposto para consecução do trabalho.

Na especificação da ontologia são detalhadas informações a respeito do domínio, propósito e público-alvo. A aquisição e extração do conhecimento serão feitos em dicionários e glossários específicos do futebol. A amostra dos termos é não probabilística, por conveniência, regida pelos critérios estabelecidos na pesquisa, que envolvem: a) termos sinônimos, polissêmicos e homógrafos; b) termos com variação diacrônica e sincrônica, além daqueles que tiveram aportuguesamento, nesses casos também serão utilizados dicionários publicados anteriormente aos acordos ortográficos de 1911; 1943 e 2009, para comparação com os termos atuais; c) termos provenientes de variedades sociolinguísticas específicas, devido aos contextos regionais. A formação do corpus de textos sobre o futebol será extraída de acervos de jornais digitais impressos. Eles servirão para validar o uso da ontologia, na recuperação da informação, de acordo com textos originais que visam comprovar a realidade.

Na conceitualização, conforme Mendonça (2015), será feita a organização e estruturação do conhecimento, por meio de artefatos de representação que vão se transformar em relações, propriedades e restrições de uma ontologia. Os artefatos propostos na OntoForInfoScience são: Dicionário de conceitos (artefato parcial); Tabela de conceitos e valores (artefato parcial); Tabela de conceitos e propriedades (final); Dicionário de verbos (final) e Modelos conceituais gráficos.

Na etapa de fundamentação ontológica, considerando que a abordagem filosófica adotada nesta pesquisa é a pragmática, serão seguidos os princípios prescritos na ontologia de fundamentação *Information Artifact Ontology* (IAO). Conforme Mendonça (2015, p. 92) esta ontologia “possibilita a representação das diferentes maneiras com as quais a informação relaciona-se com o mundo real”. Na etapa de fundamentação, o conhecimento tratado anteriormente em nível formal, passa a ser tratado em nível ontológico-formal. Para tanto, será utilizado o editor de ontologias Protegé, para criar as definições, propriedades e instâncias das classes, especificar as relações ontológicas e definir as propriedades das relações classes. Para conduzir esta etapa, o método 101 (NOY; MCGUINESS, 2001) servirá como um guia de orientação para formar exemplos práticos no editor Protegé. A etapa de avaliação será feita de acordo com conjunto de parâmetros e critérios avaliativos proposto na OntoForInfoScience.

A pesquisa no momento está na fase de definição das ferramentas de coleta automática de termos para construção do corpus. Deverá ser concluída no primeiro trimestre de 2016.

5 CONSIDERAÇÕES PARCIAIS

Os jornais são fontes de informação primárias, e em alguns casos, somente neles podemos encontrar determinadas informações, como no caso do domínio explorado, o futebol. É, portanto, importante viabilizar que usuários possam obter essas informações de maneira mais precisa e simplificada.

Considerando os aspectos linguísticos presentes nos acervos digitais, sobretudo os relativos à variação sincrônica e diacrônica da língua, espera-se que os resultados obtidos possam explorar o potencial das ontologias para operar com os processos de significação, que envolvem, dentre outros aspectos, as relações entre os termos presentes na linguagem natural. Espera-se também, destacar a utilização de acervos digitais de jornais como importante fonte de informação primária, cujo acesso deve ser facilitado por meio da utilização de ferramentas de busca orientadas para o tratamento semântico do conteúdo. Por fim, espera-se demonstrar como a modelagem de um domínio específico, nessa proposta o futebol, pode minimizar os impactos causados pela ambiguidade e variações linguísticas.

REFERÊNCIAS

ALMEIDA, M. B.; OLIVEIRA, V.; COELHO, K. Estudo exploratório sobre ontologias aplicadas a modelos de sistemas de informação: perspectivas de pesquisa em Ciência da Informação. **Encontros Bibli**: Revista Eletrônica de Biblioteconomia e Ciência da Informação, set. 2010. Disponível em:

<<http://www.periodicos.ufsc.br/index.php/eb/article/view/10987/14369>>. Acesso em: 13 set. 2015.

ALMEIDA, M. B.; MENDONÇA, F. M.; AGANETTE, E. C. Interfaces entre ontologias e conceitos seminais da Ciência da informação: em busca de avanços na organização do conhecimento. In: ENANCIB, v. 4, 2013, Florianópolis. **Anais eletrônicos...** Florianópolis, 2013. Disponível em:

<http://repositorios.questoesemrede.uff.br/repositorios/bitstream/handle/123456789/2339/INTERFACES_ENTRE_ONTOLOGIAS.pdf?sequence=1>. Acesso em: 13 set. 2015.

BERNERS-LEE, T.; HENDER, J.; LASSILA, O. **The semantic web**: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. 2001. Disponível em:

<<http://www.cs.umd.edu/~golbeck/LBSC690/SemanticWeb.html84A9809EC588EF21>>. Acesso em: 8 jul. 2015.

BREITMAN, K. K. **Web semântica**: a internet do futuro. Rio de Janeiro: LTC, 2010.

CALDAS, A. **Deu no jornal**: o jornalismo impresso na era da Internet. 2. ed. São Paulo: PUC-Rio; São Paulo: Loyola, 2004. 207 p.

DAHLBERG, I. Teoria do conceito. **Ciência da Informação**, v. 7, n. 2, p. 101-107, 1978.

DODEBEI, V. Patrimônio digital virtual: Herança, documento e informação. In: REUNIÃO BRASILEIRA DE ANTROPOLOGIA, 26., 2006, Porto Seguro, Bahia. **Anais eletrônicos...** Bahia: RBA, 2006. Disponível em: http://www.abant.org.br/conteudo/ANAIS/CD_Virtual_26_RBA/grupos_de_trabalho/trabalhos/GT%2037/vera%20dodeber.pdf. Acesso em: 8 jul. 2015.

LE GOFF, J. **História e memória**. Campinas: UNICAMP, 1990. 553 p.

MENDONÇA, Fabrício Martins. **Ontoforinfoscience**: metodologia para construção de ontologias pelos cientistas da informação: uma aplicação prática no desenvolvimento da ontologia sobre componentes do sangue humano (HEMONTA). 2015. 311 f. Tese (Doutorado)- Curso de Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

MOURA, D. O. O 'sincrônico', o 'diacrônico', o acontecimento e a errância de sentidos na análise do discurso jornalístico. **Comunicação & Informação**, Brasil, v. 2, n. 12, p. 63-73, jul. 2009. Semestral. Disponível em: <http://www.revistas.ufg.br/index.php/ci/article/view/12270/8132>. Acesso em: 15 set. 2015

NOY, N. F.; McGUINNESS, D. L. **Ontology development 101**: a guide to creating your first ontology. 2001. Disponível em: <http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html>. Acesso em: 8 jul 2015.

SOERGEL, D. **Functions of a thesaurus / classification / ontological knowledge base**. College of Library and Information Services, University of Maryland, 1997. Disponível em: <http://www.dsoergel.com/cv/soergelfctclass.pdf>. Acesso: 13 set. 2015.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: principles and methods. **Data & Knowledge Engineering**, n. 25, p. 161-167, 1998. Disponível em: <http://www.it.iitb.ac.in/~palwencha/ES/Knowledge%20engineering%20-%20Principles%20and%20methods.pdf>. Acesso em: 8 jul 2015.

VICKERY, B. C. Ontologies. **Journal of Information Science**. v. 23, n. 4. p. 227-286, jan. 1997.