Ontologias de domínio e tecnologias semânticas para promover o acesso a dados governamentais

Jaime A. Pinto, Marcelo P. Bax (professor orientador**)**

Escola de Ciência da Informação – Universidade Federal de Minas Gerais (UFMG)

Av. Antônio Carlos, 6627, Campus Pampulha, CEP 31.270-901, Belo Horizonte, Minas Gerais, Brasil

jaimepinto@eci.ufmg.br, bax@eci.ufmg.br

Abstract. *This research deals with the problem of semantic integration of heterogeneous data, focusing on Open Government Data - OGD. However, because the composition of OGD is a very broad set of data sources, with a great diversity of syntactic representation, we need to employ data integration techniques that make possible to use the set of sources, not just isolated databases. Thus, we consider the ontologies as artifacts capable of providing semantic integration, facilitating the use and dissemination of OGD content. The current state of the research aims to delimit the focus on ontologies on finance, especially those of public domain and shared construction.*

Resumo. *Esta pesquisa trata do problema de integração semântica de dados heterogêneos, com foco em Dados Governamentais Abertos - DGA. Porém, por ser o DGA representado por um conjunto muito amplo de fontes de dados, com grande diversidade de representação sintáticas, necessitamos empregar técnicas de integração de dados que possibilitem a utilização do conjunto de fontes, e não somente de arquivos isolados. Pesquisa-se a utilização de ontologias como um artefato capaz de prover a integração semântica, facilitando o uso e a disseminação do conteúdo do DGA. O estado atual da pesquisa busca delimitar o foco em ontologias sobre finanças, especialmente as de domínio público e construção compartilhada.*

1. APRESENTAÇÃO E RESUMO TEÓRICO

Os Governos são fontes de geração contínua de dados e informações, arquivando digitalmente suas atividades de planejamento, execução e controle. Estes dados, juntamente com os dados corporativos, podem compor enormes conjuntos, em formato estruturado (sob gerência de um SGBD) ou não (planilhas, textos, etc.). Estes arquivos podem ser transacionais, de alta granularidade e dispersão, ou gerenciais, centralizados em *datawarehouse* – DW. O enorme

volume produzido e acumulado constitui o ambiente que chamamos de *bigdata*, caracterizado pelos 3 V´s: Volume, Variedade e Velocidade (GANDOMI; HAIDER, 2015).

O foco deste trabalho são os Dados Governamentais Abertos – DGA – que são informações de governo disponíveis via internet, em domínio público, para uso livre pela sociedade. O fornecimento amplo de DGA tende a propiciar maior transparência e melhores oportunidades de controle social das ações governamentais (MATHEUS; RIBEIRO; VAZ, 2015; VAZ; RIBEIRO; MATHEUS, 2011).

Como há variedade de fontes de dados disponíveis, é cada vez mais difícil planejar e realizar a seleção, aquisição e combinação de dados. Os dados têm natureza heterogênea (arquivos, textos, imagens, etc.), apresentam-se em formatos diversos (SQL, XML, JPG, etc.) e tem métodos de acesso diferentes (*webservices*, API, etc.). Além disto, a interpretação destes dados da forma correta deve levar em consideração as diferenças de significado em sua origem (ALMEIDA; BAX, 2003).

O acesso aos dados, independentemente da forma de armazenamento que esteja sendo utilizada, não se restringe a leitura de dados por um usuário humano. É preciso pensar em interoperação de sistemas, ou seja, a troca de dados e a sincronização de processos informatizados. A interoperabilidade é classificada em diversos níveis: técnico, semântico, político, humano, intercomunidades e internacional (FARINELLI; MELO; ALMEIDA, 2013).

Esta pesquisa pretende ser uma contribuição da Ciência da Informação ao estudo dos problemas de acesso ao DGA utilizando ontologias de domínio e técnicas de representação semântica, com vistas à integração de dados heterogêneos e a produção de melhores aplicativos e sistemas de Governo Eletrônico – eGov.

Destaca-se a esperada aplicação prática do estudo, com a construção de um artefato de integração de alguns dos diversos arquivos DGA já publicados. Este artefato pode servir de exemplo e gerar modelos usáveis em sistemas governamentais de informação. Como software livre ou artefato de domínio público pode também fomentar o desenvolvimento de novas aplicações pela própria comunidade de desenvolvedores, facilitando ainda mais a disseminação das informações contidas no DGA.

2. PROBLEMA E OBJETIVO DE PESQUISA

No âmbito governamental torna-se cada vez mais necessária a melhoria em transparência e governança, permitindo a publicidade da execução orçamentária e a fiscalização, pelos cidadãos, dos gastos e investimentos. A simples publicação dos dados recuperados de grandes sistemas legados não é suficiente para promover o reuso e a integração. As abordagens que usam

tecnologias como XML e RDF carecem da representação do significado dos dados registrados. Abordagens mais modernas propõem o uso de ontologias (FAÇANHA; CAVALCANTI, 2014). Vale notar também o esforço empreendido pelo portal www.data.gov, cuja motivação e ideias principais estão expostas em (HENDLER *et al.*, 2012).

Parte-se da hipótese de que o problema abordado (acesso aos DGA) é candidato à solução tecnológica utilizando tecnologias semânticas. É tema relevante dadas a imposição legal e a demanda popular, em crescimento no mundo todo e no Brasil.

Os objetivos desta pesquisa são:

1. Demonstrar a viabilidade de uso das Ontologias e das técnicas semânticas para prover a integração de dados heterogêneos em ambiente de alto volume e variedade de dados.
2. Contribuir com as iniciativas de governo eletrônico provendo um artefato semântico, de domínio público, que oriente a construção de sistemas e aplicativos.

Como objetivos específicos tem-se:

1. Levantar o estado atual do esforço Dados Governamentais Abertos – DGA – do governo Brasileiro. Quais conjuntos de dados relacionados com finanças estão atualmente disponíveis?
2. Buscar formas de integração desses dados através de ontologias, baseando-se em padrões internacionais e melhores práticas.
3. Contribuir para a melhora dos mecanismos de Governo Eletrônico – eGov – através da melhor compreensão dos modelos de dados em diversos órgãos e apontando possíveis melhoras em processos e sistemas atualmente em uso.
4. Contribuir para a melhora da Governança Pública através da elaboração de mecanismos de integração, visualização e compreensão dos DGA, permitindo maior transparência aos processos e resultados de governo.
5. Projetar e disponibilizar um modelo semântico que sirva de base para a construção de ontologias de domínio específicas para aplicações de eGov.

Alguns dos ganhos gerais desta orientação ontológica são relacionados na lista abaixo, de maneira simplificada (FREITAS, 2003):

- A possibilidade de reuso de ontologias já prontas e bases de conhecimento públicas, fazendo adaptações e extensões.
- A existência de um razoável número de ontologias públicas (ditas “de prateleira”) disponíveis para uso, consulta e adaptações. No caso de finanças (possível escopo da

pesquisa), esta possibilidade parece especialmente valiosa pela possibilidade de reuso intensional de conhecimento de outras áreas, de ciências exatas e sociais.

- O acesso on-line a servidores de ontologias que, armazenando milhares de classes e instâncias, podem funcionar como mantenedores de integridade do conhecimento compartilhado, buscando garantir a uniformidade do vocabulário.
- A possibilidade de integração e interoperação de bases de dados já existentes e de sistemas legados através do mapeamento entre formalismos de representação do conhecimento. Isto pode disponibilizar uma enorme massa de dados armazenados hoje em bancos de dados relacionais, através de uma interface de acesso comum.

3. METODOLOGIA E ESTÁGIO ATUAL

O trabalho em andamento é a realização uma pesquisa aplicada – aplicação dos conhecimentos básicos na geração de novos produtos, processos e serviços – com objetivos exploratórios – descoberta de teorias e práticas que modificarão as existentes – empregando procedimentos de pesquisa experimental – descoberta de novos materiais, métodos, técnicas, protótipos de software (JUNG, 2004).

Pretende-se aplicar neste trabalho a metodologia de projeto *Design Science Research* – DSR – conforme proposto em (BAX, 2014). Esta abordagem parece ser especialmente adequada para este trabalho, que tem como um dos objetivos a construção de um artefato: uma ontologia ou um modelo semântico para DGA.

O trabalho iniciou-se em agosto de 2016 e tem término previsto para julho de 2020. No primeiro semestre de 2017 – data atual – o projeto está na fase de Preparação da pesquisa, executando uma Revisão Sistemática da Literatura – SLR – buscando a necessária delimitação do escopo, com a escolha de um domínio de trabalho. Há uma tendência para foco no domínio financeiro, onde encontramos significativas oportunidades de realização de nossos objetivos.

Os próximos passos são:

- Identificar um subdomínio dos dados (DGA) em finanças;
- Identificar possíveis ontologias pré-existentes para conceituar esse tema;
- Fazer um experimento de integração, utilizando técnicas de *ontology matching* (OTERO-CERDEIRA; RODRIGUEZ-MARTINEZ; GOMEZ-RODRIGUEZ, 2015).
- Construir aplicações que sirvam de modelo para desenvolvimentos futuros, como por exemplo, geração de grafos de conhecimento (*knowledge graphs*), apresentação interativa de *dashboards*, agentes inteligentes para busca semântica, pesquisa por

linguagem natural e conversacional (*chatbot*) (AUGELLO *et al.*, 2009; GARCÍA-SÁNCHEZ *et al.*, 2011; SANTOS *et al.*, 2017).

4. OBSERVAÇÕES FINAIS

Este trabalho destina-se à apresentação no Workshop de Teses e Dissertações em Ontologias que será realizado em conjunto com o ONTOBRÁS 2017 – Seminário de Pesquisas em Ontologias do Brasil – no período 28-30 agosto de 2017, em Brasília. Apresenta-se a pesquisa em andamento no Programa de Pós-graduação em Gestão & Organização do Conhecimento – PPGGOC – da Escola de Ciência da Informação – ECI – da UFMG.

No estágio atual da pesquisa busca-se a fundamentação teórica do assunto, através de uma Revisão Sistemática da Literatura. Para fins de delimitação do escopo o foco está em ontologia sobre finanças, especialmente as ontologias construídas colaborativamente e de domínio público.

Espera-se poder beneficiar-se de esforços mundiais já em andamento como, por exemplo, a FIBO™ - *Financial Industry Business Ontology*. FIBO™ é uma ontologia financeira que começou a ser desenvolvida em 2008 e que ainda está em desenvolvimento aberto e colaborativo, através do esforço conjunto de diversas entidades de âmbito mundial. Seu objetivo é construir um padrão de ampla aceitação para os negócios da indústria financeira. Alguns de seus patrocinadores mais importantes são ISO, W3C e EDMC (BENNETT, 2013, 2014; LOEHRLEIN; LEMIEUX; BENNETT, 2014).

BIBLIOGRAFIA

ALMEIDA, Maurício Barcellos. *Inter-operabilidade entre fontes heterogêneas: um meta-modelo baseado em ontologias*. 2002. 146 f. UFMG, 2002. Disponível em: <http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/LHLS-6AZQHU/mestrado___mauricio_barcellos_almeida.pdf?sequence=1>. Acesso em: 28 jun. 2016.

ALMEIDA, Maurício Barcellos; BAX, Marcello Peixoto. Taxonomia para projetos de integração de fontes de dados baseados em ontologias. 2003, Belo Horizonte: [s.n.], 2003. p. 1–20. Disponível em: <http://mba.eci.ufmg.br/downloads/artigo_taxonto_sub.pdf>.

AUGELLO, A *et al.* A Semantic Layer on Semi-Structured Data Sources for Intuitive Chatbots. 2009, [S.l: s.n.], 2009. p. 760–765.

BAX, Marcello Peixoto. Design Science: Filosofia Da Pesquisa Em Ciência Da Informação E Tecnologia. *XV Encontro Nacional de Pesquisa em Ciência da Informação – ENANCIB 2014*, n. XV ENANCIB-AL{É}M DAS NUVENS, p. 3883–3903, 2014.

BENNETT, Michael. Adopting and Extending REA Terms in the Financial Industry Business Ontology : A Case Study. 2014, Berlin: [s.n.], 2014. p. 1–5.

BENNETT, Michael. The financial industry business ontology: Best practice for big data. *Journal of Banking Regulation*, v. 14, n. 3–4, p. 255–268, 2013. Disponível em: <http://www.palgrave-journals.com/jbr/journal/v14/n3/abs/jbr201313a.html>.

FAÇANHA, Raquel Lima; CAVALCANTI, Maria Cláudia. On the Road to Bring Government Legacy Systems Data Schemas to Public Access. 2014, Rio de Janeiro: CEUR Workshop Proceedings, 2014.

FARINELLI, Fernanda; MELO, Stefane; ALMEIDA, Maurício Barcellos. O papel das ontologias na interoperabilidade de sistemas de informação: reflexões na esfera governamental. *XIV Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB 2013)*, p. 1–21, 2013. Disponível em: <http://mba.eci.ufmg.br/downloads/Interoperab_Enancib_2013_camera-ready.pdf>.

FREITAS, Fred. Ontologias e a Web Semântica. *Anais do XXIII Congresso da Sociedade Brasileira de Computação*, p. 1–52, 2003. Disponível em: <http://www.inf.ufsc.br/~fernando.gauthier/EGC6006/material/Aula 3/Ontologia_Web_semantica Freitas.pdf>.

GANDOMI, Amir; HAIDER, Murtaza. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, v. 35, n. 2, p. 137–144, 2015.

GARCÍA-SÁNCHEZ, Francisco *et al.* Applying intelligent agents and semantic web services in eGovernment environments. *Expert Systems*, v. 28, n. 5, p. 416–436, 2011. Disponível em: <http:https://dx.doi.org/10.1111/j.1468-0394.2011.00586.x>.

HENDLER, James *et al.* US government linked open data: Semantic.data.gov. *IEEE Intelligent Systems*, v. 27, n. 3, p. 25–31, 2012.

JUNG, Carlos Fernando. *Metodologia para pesquisa & desenvolvimento: aplicada a novas tecnologias, produtos e processos*. 1. ed. São Paulo: [s.n.], 2004.

LOEHRLEIN, Aaron J.; LEMIEUX, Victoria L.; BENNETT, Michael. The classification of financial products. *Journal of the Association for Information Science and Technology*, v. 65, n. 2, p. 263–280, 18 fev. 2014. Disponível em: <http://doi.wiley.com/10.1002/asi.22969>. Acesso em: 15 out. 2015.

MATHEUS, Ricardo; RIBEIRO, Manuella Maia; VAZ, José Carlos. Brazil towards government 2.0: Strategies for adopting open government data in national and subnational governments. *Case Studies in E-Government 2.0: Changing Citizen Relationships*, v. 55, n. 11, p. 121–138, 2015. Disponível em: <http://vaz.blog.br/blog/wp-content/uploads/2015/02/11-Brazil-Toward-smart-and-participative-government.pdf>.

OTERO-CERDEIRA, Lorena; RODRIGUEZ-MARTINEZ, Francisco J.; GOMEZ-RODRIGUEZ, Alma. Ontology matching: A literature review. *Expert Systems with Applications*, v. 42, n. 2, p. 949–971, 2015.

SANTOS, Henrique *et al.* From Data to City Indicators: A Knowledge Graph for Supporting Automatic Generation of Dashboards. 2017.

VAZ, José Carlos; RIBEIRO, Manuella Maia; MATHEUS, Ricardo. Dados governamentais abertos e seus impactos sobre os conceitos e práticas de transparência no Brasil. *Cadernos PPG-AU/UFBA*, v. 9, n. 1, p. 45–62, 2011. Disponível em: <http://www.portalseer.ufba.br/index.php/ppgau/article/view/5111>.