

Modelo Conceitual de Dados Baseado em Ontologia: Estudo de Caso CGU

¹Márcia Myuki Takenaka Fujimmoto and ²Edna Dias Canedo

¹ Ministério da Transparência e Controladoria-Geral da União - CGU

² Mestrado Profissional em Computação Aplicada, Departamento de Ciência da Computação
Universidade de Brasília - (UnB) – Brasília-DF, Brasil

marcia.takenaka@cgu.gov.br, ednacanedo@unb.br

Abstract. *This paper describes the study and proposal of an information architecture, in the form of an ontology-based conceptual data model, for the CGU Ministry with potential for use in the implementation of data sources governance and management based on Cobit 4.1 PO2 - Define the Information Architecture process.*

keywords: *Data Management, Ontology, Integration, Information Architecture*

Resumo. *Este artigo descreve o estudo e a proposta de arquitetura da informação, sob a forma de um modelo conceitual de dados baseado em ontologia, para o Ministério CGU com potencial de uso na implementação da governança e gestão de fontes de dados baseada no processo PO2 - Definir a Arquitetura da Informação do Cobit 4.1.*

1. Introdução

O conceito de ontologia tem recebido atenção nos últimos anos, particularmente como abordagem para a arquitetura e gestão da informação, para a web semântica e para a mineração e integração de dados. A aplicabilidade variada da ontologia direcionou as investigações deste trabalho no sentido de considerá-la como base para uma proposta de arquitetura da informação (AI) de suporte à gestão de dados organizacionais do Ministério da Transparência e Controladoria-Geral da União - CGU.

As atuação da CGU é diversificada [CGU 2017] e envolve: disponibilização de informações em aplicações web; cruzamento de dados para identificar ilícitos; mineração de dados em busca de indícios de irregularidades. As fontes de dados utilizadas são numerosas e diversificadas, há mais de 6.800 tabelas de banco de dados (BD), muitas provenientes de sistemas estruturantes do Governo Federal. A utilização eficaz e eficiente dessas fontes envolve o reúso de dados e a interoperabilidade entre sistemas, o que requer uma governança e gestão de dados bem estruturada, no entanto, a estrutura de suporte à gestão de dados existente no Órgão é elementar.

A preocupação com o aprimoramento da governança de dados fez com que a CGU iniciasse planos para a implantação do processo PO2 - Definir a Arquitetura da Informação do Cobit 4.1 [ISACA 2007]. Os objetivos de controle desse processo atentam para a necessidade do entendimento comum dos dados entre a Tecnologia da Informação

e o Negócio; do compartilhamento dos elementos de dados entre aplicativos e sistemas; da prevenção quanto à criação de elementos de dados incompatíveis. No decorrer da elaboração dos planos, foi elaborado um modelo de AI no estilo Entidade-Relacionamento (ER), chamado de Meta-Modelo CGU, cujo foco foram metadados sob o ponto de vista de BD. Esse modelo, ainda não implementado, carece de elementos conceituais, que auxiliem o entendimento dos dados, e não apresenta estrutura adequada de suporte ao compartilhamento de dados. Assim, este trabalho, sob o ponto de vista de aplicação ao caso CGU, faz uma proposição de modelo de AI sob a forma de modelo conceitual de dados baseado em ontologia como forma de contribuir para o alcance de objetivos de controle de PO2. Por outro lado, sob o ponto de vista da academia, este trabalho pode demonstrar a aplicabilidade da abordagem ontológica de gestão de dados e de arquitetura de referência para estrutura ontológica propostas em estudos anteriores.

Este artigo está organizado como segue. A Seção 2 apresenta algumas definições consideradas importantes para o entendimento deste trabalho. A Seção 3 relata alguns resultados de mapeamento sistemático efetuado e alguns trabalhos correlatos. As Seções 4 e 5 apresentam a proposta de AI para a CGU e resultados da sua validação, respectivamente. Por fim, a Seção 6 apresenta as conclusões e os trabalhos futuros.

2. Fundamentação

De acordo com Borst [Borst 1997], uma ontologia é definida como uma **especificação formal e explícita** de uma **conceitualização compartilhada**, na qual a especificação formal quer dizer algo que é legível para os computadores; explícita são os conceitos, propriedades, relações, funções, restrições e axiomas explicitamente definidos; conceitualização representa um modelo abstrato de algum fenômeno do mundo real e compartilhada significa conhecimento consensual.

As linguagens usadas na especificação de modelos conceituais suportados por Ontologia, como UML *Unified Modeling Language*, ER e ORM *Role Modeling* são frequentemente linguagens orientadas a software. Como consequência, a falta de semântica nos modelos criados com essas linguagens leva a problemas como interpretações equivocadas, representações ambíguas e imprecisas e dificuldades na comunicação [Albuquerque 2013]. Para remediar essa situação, na última década surgiu o que é conhecida como modelagem conceitual suportada por ontologia, que consiste na prática de usar ontologias e princípios ontológicos para apoiar a atividade de modelagem conceitual [Burton-Jones and Meso 2002], [Parsons and Cole 2004], [Evermann and Wand 2005], [Li and Parsons 2007].

2.1. Integração de Dados

A definição de integração de dados adotada neste trabalho é a de [Lenzerini 2002]: é a capacidade computacional de abordar o problema de fornecer dados localizados em diversas fontes heterogêneas sob uma perspectiva única. Segundo [Calvanese et al. 1998], há basicamente duas abordagens para se revolver o problema de integração de dados: a procedural e a declarativa. O objetivo da abordagem declarativa, de interesse deste trabalho, é modelar os dados nas fontes utilizando uma linguagem adequada a fim de construir uma representação unificada dos dados a ser referenciada para pesquisas futuras. A integração e a unificação de dados transversais a várias aplicações foram as duas principais formas de compartilhamento de dados consideradas nestes estudos.

3. Revisão de Literatura

A primeira etapa deste trabalho foi um mapeamento sistemático, que permitiu a construção de um panorama da atenção recebida pelos estudos que assentaram suas ideias na utilização da ontologia para a gestão ou integração de dados. A distribuição por ano de 252 trabalhos selecionados em busca refinada produziu uma linha de tendência crescente para o período entre 2006 e 2107, demonstrando um crescimento nas pesquisas relacionadas ao tema.

A partir de questões de pesquisa formuladas e respondidas com base em 28 artigos selecionados para leitura integral foi possível observar que: **i.** o foco dos estudos concentrou-se em propostas de modelos empíricos (22 artigos); **ii.** há 15 estudos que continham proposta de ontologia juntamente com a solução técnica na qual a ela será utilizada; **iii.** muitos dos estudos relatam necessidade de ampliação de uso ou evoluções de suas propostas. Há número reduzido de estudos (16) que discorreram sobre a implementação/avaliações de suas propostas; **iv.** os tópicos mais mencionados como pesquisa futura foram: *matching* de ontologias, buscas semânticas e anotações semânticas.

3.1. Trabalhos Relacionados

A ideia básica da Abordagem Ontológica de Gestão de Dados (OBDM) apresentada por [Lenzerini 2011] é uma arquitetura em 3 camadas: **1.** Uma ontologia do domínio; **2.** Uma camada das fontes de dados e **3.** Uma camada de mapeamento entre as duas anteriores (1 e 2). Essa separação apresenta algumas vantagens potenciais [Daraio et al. 2016]: **i.** a camada ontológica é o meio que leva a uma abordagem integrativa na governança de dados; **ii.** a ontologia e os mapeamentos correspondentes das fontes de dados proveem uma base para a documentação de todos os dados organizacionais ; **iii.** extensibilidade do sistema, ou seja, a partir de um esqueleto básico do domínio pode se adicionar incrementalmente novas fontes de dados ou novos elementos.

A arquitetura de referência para estrutura ontológica proposta por [Fitzpatrick et al. 2012] é composta por: **i.** Ontologias de Fundamentação (ou Ontologias de Alto Nível), **ii.** Ontologia de Integração de Dados Multidomínio, **iii.** Ontologias de Domínio e **iv.** Ontologias de Tarefas e **v.** Ontologia de Aplicação. [Fitzpatrick et al. 2012] explica que à luz do critério da ortogonalidade, a ontologia de integração de dados multidomínio é relacionada, por subsunção, a ontologias de fundamentação como a SUMO e outras. As ontologias de domínio são incluídas na ontologia de dados de multidomínio. A estrutura compreende, também, ontologias de tarefas genéricas e ontologia de aplicações, esta última para suportar tarefas específicas de domínio.

4. Proposta

No âmbito deste estudo, não foi possível construir uma conceitualização compartilhada por todos os interessados, por isso, optou-se por adotar um modelo conceitual baseado em ontologia como AI proposta. O desenvolvimento do modelo seguiu uma customização da abordagem "*Enterprise*" limitada à sua fase informal (composta por identificação do propósito e identificação do escopo) [Jones et al. 1998]. Tal delimitação não prejudica a criação do esqueleto do modelo: estrutura e os respectivos conceitos-chave ou abrangentes, estando coerente com a extensibilidade prevista na abordagem OBDM. O propósito

estabelecido para o modelo foi: "Estruturar a organização dos metadados das fontes de dados CGU de forma a permitir o entendimento abrangente das fontes e o seu compartilhamento, viabilizando o reúso de dados e a interoperabilidade entre sistemas.". Quanto ao nível de formalidade, optou-se por um modelo equivalente a uma ontologia "Estruturada Informal"[Uschold and Gruninger 1996], suficiente quando não há processamento do modelo diretamente por máquina. As atividades de identificação do escopo, na qual é produzida a especificação do domínio a representar, e de avaliação do modelo se utilizaram de questões de competência informais.

O modelo holístico proposto é baseado na arquitetura de referência de estrutura ontológica [Fitzpatrick et al. 2012] e está segmentado nas camadas de: **i.** Dados Multidomínio - MD; **ii.** Domínios de Negócio - DN; **iii.** Tarefas-Gerais - TG; **iv.** Aplicações e Sistemas - AP. A estrutura formada pelas camadas acrescidas dos respectivos conceitos-chave é a base da hierarquia formada no protótipo desenvolvido na ferramenta Web-Protégé [Protégé]. O protótipo inclui relacionamentos e definições de forma a completar a representação da camada **1**, ontologia, da OBDM.

4.1. Camadas do Modelo de Arquitetura da Informação

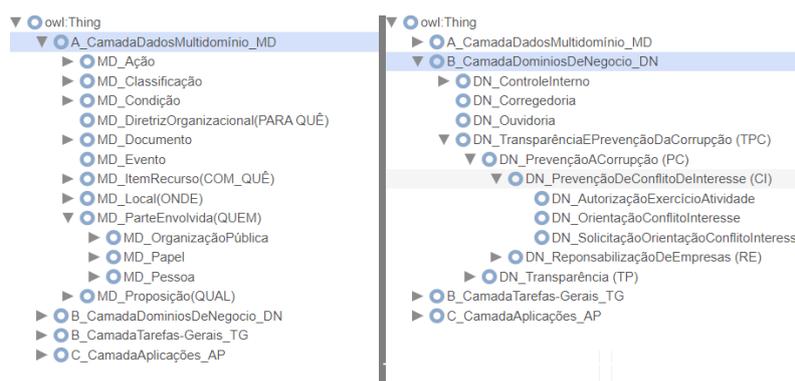


Figura 1. Classes da Camada Integração MD e de Domínio de Negócios DN

A principal finalidade da camada MD é ser abrangente o suficiente para conter todos os conceitos fundamentais e transversais aos vários domínios de negócio CGU. O lado esquerdo da Figura 1 identifica as classes que compõem essa camada, que foram selecionadas principalmente de ontologias organizacionais utilizadas como referência e acrescidas de determinados conceitos gerais.

A camada DN tem como finalidades: prover uma visão das fontes de dados a partir dos domínios de negócio CGU e explicar a essência de cada um desses domínios a partir da descrição dos relacionamentos entre classes. As classes no topo da hierarquia formada DN representam as competências da CGU [CGU 2017]. Os domínios de negócio da CGU são explicados por relacionamentos que expressem a seguinte forma geral: Sujeito (Ex:Partes Envolvidas) + Ação (Ex:Tipo de Ação do Órgão) + Objeto (Ex:Item de Recurso) + Complementos (Ex:Local ou Documento).

A finalidade da camada TG é revelar os serviços ou tarefas de suporte que são independentes de domínio de negócio. Os assuntos Arquivo, Segurança, E-mail, *Workflow* são exemplos de classes dessa camada. Por último, a camada de aplicações AP tem como principal finalidade mostrar a lista dos sistemas/aplicações CGU, classes principais

da camada, com subclasses representativas das fontes de dados que as compõem. O relacionamento de cada sistema/aplicação com suas subclasses (que podem ser provenientes de MD, DN, TG ou da própria camada AP) é de composição "Is part of".

A utilização do modelo produzido pode ser numa abordagem *top-down*, *bottom-up* ou por camada específica, a depender da necessidade de compartilhamento ou entendimento das fontes de dados. Por exemplo, quando a intenção é integrar ou unificar de bases de dados, as classes MD são o principal alvo de pesquisa. Por outro lado, quando há necessidade de produção de informações estratégicas sobre as atividades do órgão, o entendimento dos domínios de negócio CGU e das fontes subjacentes pode ocorrer partindo-se das classes DN e, em seguida, examinando os seus relacionamentos com classes das outras camadas.

5. Validação do Modelo Proposto

O protótipo construído foi verificado primeiramente a partir do mapeamento de tabelas em BD da CGU. Para isso foram criados na WebProtégé elementos correspondentes à camada 2 da OBDM, metadados de 83 (oitenta e três) tabelas de BD, cobrindo as principais entidades existentes no Meta-Modelo CGU. Em seguida, foram registrados os mapeamentos entre tabelas e classes do modelo, que correspondem à camada 3 da OBDM, considerando as tabelas como individuais (instâncias) da classe para a qual deve ser mapeada. Para efeito de simplificação, não houve mapeamento no nível de atributo e as tabelas foram mapeadas somente para o seu principal conceito ou classe. Foi possível efetuar o mapeamento de todas as tabelas, porém alguns implicaram em criação de novas subclasses para melhor adequar o relacionamento fonte-classe.

A validação do protótipo foi completada por avaliação de dez especialistas do Órgão por meio de questionário com duas seções para: **1.** verificar se o modelo atende às questões de competência que guiaram a sua elaboração; **2.** verificar a percepção de melhorias da AI proposta neste trabalho com relação ao Meta-Modelo. Setenta e seis por cento das manifestações quanto às perguntas da primeira seção foram de que o modelo retorna respostas completas às questões de competência formuladas. As respostas à segunda seção permitiram concluir que a percepção dos analistas é de que o modelo é capaz de agregar melhorias importantes ao Meta-Modelo CGU (oitenta por cento das manifestações às perguntas da segunda seção).

6. Conclusão e Trabalhos Futuros

Os resultados obtidos com a validação do modelo de AI deste trabalho foram considerados satisfatórios e permitem concluir que a proposta de AI deste trabalho é factível e pode auxiliar o atendimento dos objetivos de controle do processo Cobit PO2- Definir a Arquitetura da Informação na CGU, notadamente no que diz respeito ao entendimento e ao compartilhamento de dados.

Como trabalhos futuros, há interesse nas seguintes atividades: **1.** aprimorar o esqueleto de AI proposto, tornando-o um modelo mais completo e formal; **2.** estudar as técnicas existentes para a conversão de modelos ontológicos em modelos relacionais; **3.** identificar e avaliar ferramentas de correspondência automática de estruturas de dados a fim de auxiliar a atividade de mapeamento entre fontes e classes.

Referências

- Albuquerque, A. F. (2013). Ontological foundations for conceptual modeling datatypes.
- Borst, W. N. (1997). *Construction of engineering ontologies for knowledge sharing and reuse*. Universiteit Twente.
- Burton-Jones, A. and Meso, P. (2002). How good are these uml diagrams? an empirical test of the wand and weber good decomposition model. *ICIS 2002 Proceedings*, page 10.
- Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D., and Rosati, R. (1998). Description logic framework for information integration. In *KR*, pages 2–13.
- CGU (2017). <http://www.cgu.gov.br/sobre/institucional/competencias-e-organograma>. Acessado: 27 ago. 2017.
- Daraio, C., Lenzerini, M., Leporelli, C., Moed, H. F., Naggar, P., Bonaccorsi, A., and Bartolucci, A. (2016). Data integration for research and innovation policy: an ontology-based data management approach. *Scientometrics*, 106(2):857–871.
- Evermann, J. and Wand, Y. (2005). Toward formalizing domain modeling semantics in language syntax. *IEEE Transactions on Software Engineering*, 31(1):21–37.
- Fitzpatrick, D., Coallier, F., and Ratte, S. (2012). *A Holistic Approach for the Architecture and Design of an Ontology-Based Data Integration Capability in Product Master Data Management*, pages 559–568. Springer Berlin Heidelberg, Berlin, Heidelberg.
- ISACA (2007). Control objectives for information and related technologies - cobit 4.1. <http://www.isaca.org/Knowledge-Center/cobit/Pages/Downloads.aspx>. Acessado: 27 ago. 2017.
- Jones, D., Bench-Capon, T., and Visser, P. (1998). Methodologies for ontology development.
- Lenzerini, M. (2002). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246. ACM.
- Lenzerini, M. (2011). Ontology-based data management. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM '11*, pages 5–6, New York, NY, USA. ACM.
- Li, X. and Parsons, J. (2007). Ontological semantics for the use of uml in conceptual modeling. In *Tutorials, posters, panels and industrial contributions at the 26th international conference on Conceptual modeling-Volume 83*, pages 179–184. Australian Computer Society, Inc.
- Parsons, J. and Cole, L. (2004). An experimental examination of property precedence in conceptual modelling. In *Proceedings of the first Asian-Pacific conference on Conceptual modelling-Volume 31*, pages 101–110. Australian Computer Society, Inc.
- Protégé, T. S. U. Webprotégé. <https://webprotege.stanford.edu/>. Acessado: 05 jun. 2018.
- Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(2):93–136.