Integração de Bancos de Dados Tributários: um Estudo de Caso Aplicado à Prefeitura de São Paulo

André Ippolito[1,2] , Edit Grassiani Lino de Campos[1]

[1]Instituto de Pesquisas Tecnológicas do Estado de São Paulo (IPT)
Av. Prof. Almeida Prado, 532 - Cid. Universitária - CEP 05508-901 - São Paulo – SP – Brazil

[2]Prefeitura do Município de São Paulo – Secretaria de Finanças
Viaduto do Chá, n.º 15 – Centro – CEP 01002-020 – São Paulo – SP - Brazil

aippolito@prefeitura.sp.gov.br, edit.campos@gmail.com

Abstract. *Heterogeneities in database schemas normally occur because they are developed by different staffs, which follow different patterns, making it difficult to obtain alignments. This article aims to highlight the contribution of schemas alignment in the integration of taxes databases. This contribution was verified in a case study based on a tax administration application from the municipality of São Paulo. The integration of taxes databases allows the development of a system that provides integrated tax information.*

Resumo. *Heterogeneidades em esquemas de bancos de dados normalmente ocorrem, pois são desenvolvidos por diferentes equipes, que seguem padrões diversos, dificultando a obtenção de alinhamentos. Este artigo visa ressaltar a contribuição do alinhamento na integração de bancos de dados tributários. Essa contribuição verificou-se em um estudo de caso baseado em uma aplicação da administração tributária da Prefeitura de São Paulo. A integração dos bancos de dados tributários permite o desenvolvimento de um sistema que forneça informações tributárias integradas.*

1. Introdução

De acordo com [Euzenat e Shvaiko 2010], alinhamento é o conjunto de correspondências entre elementos de duas ou mais ontologias. Alinhamento é definido por [Euzenat e Shvaiko 2010] de forma mais abrangente em relação a ontologias, podendo ser aplicado também a esquemas de bancos de dados. Com a crescente quantidade de informações, dispersas em diversos bancos de dados, obedecendo a diferentes padrões e convenções, torna-se cada vez mais árdua a tarefa de obter alinhamentos. Observam-se diferentes nomes para uma mesma entidade, diferentes tipos de dados e nomes para mesmos atributos, dentre outras heterogeneidades. Conforme [Euzenat e Shvaiko 2010], a obtenção do alinhamento é o primeiro passo de ordem técnica no processo de integração de dados.

O problema de gerenciar heterogeneidades para integrar dados ocorre na administração tributária da Prefeitura de São Paulo. Seus bancos de dados foram concebidos em diferentes épocas, conforme diferentes padrões, o que gerou heterogeneidades nos esquemas dos bancos de dados. Como se trata de muitas heterogeneidades entre os esquemas, a obtenção manual do alinhamento atrasa o processo de integração. A administração tributária necessita integrar seus bancos de

dados para que tenha uma visão integrada dos contribuintes. Essa visão permite identificar os maiores devedores, possibilitando direcionar ações de orientação e cobrança a contribuintes com maior inadimplência. Isto permite efetuar um planejamento tributário mais eficiente, com maior potencial de retorno em termos arrecadatórios.

Foi efetuado um estudo de caso de alinhamento de esquemas de bancos de dados tributários heterogêneos, pertencentes à Prefeitura de São Paulo. Aplicou-se a ferramenta de alinhamento COMA [Hai 2005] [Massmann et al. 2011], na sua versão 3.0 *Community Edition* (CE) [COMA 3.0 - Community Edition 2012]. A versão da COMA 3.0 CE foi obtida em 11/06/2012. Mais publicações sobre a COMA encontram-se disponíveis em [Schema and Ontology Matching with Coma 3.0 2012].

O presente trabalho visa apresentar os resultados e contribuições obtidos em um estudo de caso baseado em um cenário real de uma administração tributária municipal. O alinhamento obtido no estudo de caso contribui para o processo de integração dos bancos de dados tributários da Prefeitura de São Paulo, pois a obtenção do alinhamento é a etapa inicial da integração. O uso de uma base de conhecimento, estruturada por meio de uma ontologia, contribuiu para o alinhamento de heterogeneidades de atributos compostos e de hiperonímia, não detectadas pela COMA 3.0 CE no estudo de caso. O restante do artigo estrutura-se da seguinte forma: na seção 2, são explicados os fundamentos teóricos referentes a heterogeneidades de esquemas de bancos de dados e técnicas de alinhamento; na seção 3, são apresentados os principais trabalhos recentes relativos a ferramentas de alinhamento, comparando-se as ferramentas conforme as técnicas que utilizam; na seção 4, é descrito o estudo de caso e seus resultados; na seção 5, são apresentadas conclusões e sugestões para trabalhos futuros.

2. Fundamentos teóricos

Nesta seção, são explicadas as principais heterogeneidades de esquemas de bancos de dados. São explicadas também técnicas de alinhamento, aplicáveis a esquemas de bancos de dados e a ontologias.

2.1. Principais heterogeneidades de esquemas de bancos de dados

Conforme [Kim e Seo 1991], as principais heterogeneidades de esquemas de bancos de dados referem-se a nomes, estruturas e domínios.

Heterogeneidades de nomes são relativas à nomenclatura dos elementos dos esquemas em comparação. São subdivididas em: sinonímia [Kim e Seo 1991], quando nomes diferentes são utilizados para o mesmo elemento (nomes diferentes com mesmo significado); homonímia [Kim e Seo 1991], quando nomes idênticos ou semelhantes referem-se a elementos diferentes (nomes iguais com significados diferentes); hiperonímia [Miller 1993], quando o conceito associado ao nome de um elemento é mais genérico que o conceito do nome de outro elemento (“árvore” é um hiperônimo de “conífera”, por exemplo).

Heterogeneidades de estruturas [Kim e Seo 1991] ocorrem quando há diferentes construções de modelagem ou diferentes restrições de integridade para um mesmo elemento de um esquema. Podem ser, por exemplo, referentes a atributos compostos. Nesses casos, em um esquema o atributo é simples (um único campo em uma tabela) e,

no outro esquema, seu similar é representado por um conjunto de atributos (dois ou mais campos de uma tabela).

Heterogeneidades de domínios [Kim e Seo 1991] ocorrem quando elementos de esquemas têm diferentes representações e precisões. Podem referir-se a elementos de esquemas com diferentes tipos de dados, unidades de medidas, precisões ou faixas de valores.

2.2. Técnicas de alinhamento

Técnicas de alinhamento visam obter um conjunto de correspondências entre elementos de ontologias e aplicam-se também a esquemas de bancos de dados. Dentre as técnicas de alinhamento, têm-se técnicas baseadas em: cadeias de caracteres, métodos intrínsecos, restrições, dicionários, reuso de resultados, dentre outras. Técnicas baseadas em cadeias de caracteres consideram nomes dos elementos como sequências de caracteres. Duas sequências são comparadas, por exemplo, por meio de funções numéricas, tais como a distância de Hamming [Hamming 1950], que calcula a similaridade entre sequências considerando o número de caracteres diferentes em duas cadeias em comparação. Técnicas baseadas em métodos intrínsecos [Euzenat e Shvaiko 2010] utilizam, por exemplo, espaços, pontuações e traços para quebrar sequências de caracteres em *tokens* e depois aplicam técnicas de cadeias de caracteres aos *tokens*. Técnicas baseadas em restrições [Euzenat e Shvaiko 2010] consideram tipos de dados, chaves e cardinalidades. Reuso de resultados de alinhamento [Euzenat e Shvaiko 2010] reaproveita correspondências existentes em novas comparações entre esquemas ou ontologias. Para um maior detalhamento a respeito de técnicas de alinhamento recomenda-se consultar [Euzenat e Shvaiko 2010].

3. Ferramentas de alinhamento

Diversas ferramentas foram desenvolvidas para obter o alinhamento entre ontologias e esquemas de bancos de dados. Dentre as mais recentes, destacam-se: COMA 3.0 CE [COMA 3.0 - Community Edition 2012]; OWL Lite Aligner [Euzenat e Valtchev 2004]; S-Match [Shvaiko et al. 2009]; Naive Ontology Mapping [Ehrig e Sure 2004]; Quick Ontology Mapping [Ehrig e Staab 2004]; Artemis [Castano et al. 2001]; Cupid [Madhavan et al. 2001]; Anchor-PROMPT [Noy e Musen 2001]. Aplicando-se a classificação de técnicas de alinhamento proposta em [Euzenat e Shvaiko 2010] às ferramentas, obtém-se o quadro comparativo apresentado na tabela 1.

Da análise da tabela 1, verifica-se que a COMA 3.0 CE é a mais abrangente em termos de técnicas de alinhamento utilizadas, sendo, portanto, escolhida para ser usada no estudo de caso.

4. Estudo de caso

O estudo de caso baseou-se em uma aplicação pertencente à administração tributária da Prefeitura de São Paulo. Trata-se de bancos de dados referentes a cadastro de contribuintes, parcelamentos, notas fiscais eletrônicas, declarações eletrônicas, pagamentos de tributos, autos de infrações, dívida ativa e gestão de expedientes. Foram comparados visualmente pares de esquemas, considerando-se todos os pares possíveis. Para cada par de esquemas, foram feitas as seguintes atividades: identificação visual das

Tabela 1. Tabela comparativa das ferramentas considerando as técnicas de alinhamento utilizadas

Ferramenta	Técnicas											Total de técnicas
	Caracteres	Métodos intrínsecos	Restrições	Dicionários	Reuso	Ontologia como base de conhecimento	Estatística	Grafo	Taxonomia	Repositório de estruturas	Modelos baseados em lógica de descrição	
COMA 3.0 CE	X	X	X	X	X	-	X	X	-	X	-	8
OWL Lite Aligner	X	X	X	X	-	-	-	X	X	-	-	6
S-Match	X	X	-	X	-	-	-	X	-	-	X	5
Naive Ontology Mapping	X	-	X	X	-	-	-	X	X	-	-	5
Quick Ontology Mapping	X	-	X	X	-	-	-	X	X	-	-	5
Artemis	-	X	X	X	-	-	-	X	-	-	-	4
Cupid	X	X	-	X	-	-	-	X	-	-	-	4
Anchor-PROMPT	X	-	X	-	-	-	-	X	-	-	-	3

heterogeneidades; classificação das heterogeneidades conforme [Euzenat e Shvaiko 2010]; aplicação das técnicas da COMA 3.0 CE; anotação das heterogeneidades detectadas e não detectadas pela ferramenta. Essas atividades foram efetuadas por auditores fiscais da Secretaria de Finanças da Prefeitura de São Paulo. Para heterogeneidades de nome, foram usados algoritmos da COMA 3.0 CE que aplicam técnicas baseadas em cadeias de caracteres e dicionários. Para heterogeneidades de tipos de dados, foram usados algoritmos da ferramenta que possuem técnicas baseadas em restrições. Para heterogeneidades de hiperonímia e de atributos compostos, foram implementados algoritmos adicionais, uma vez que a COMA 3.0 CE não tem técnicas para essas heterogeneidades.

4.1. Heterogeneidades existentes na aplicação

Da análise visual dos esquemas da aplicação, resultaram 207 heterogeneidades: 153 heterogeneidades de nome (117 de sinonímia e 36 de hiperonímia); 34 de tipos de dados e 20 estruturais de atributos compostos.

Na tabela 2, são apresentados alguns exemplos de heterogeneidades encontradas no estudo de caso.

4.2. Resultados com a utilização dos algoritmos da COMA 3.0 CE

Do total de 207 heterogeneidades do estudo de caso, a COMA 3.0 CE detectou todas as heterogeneidades de sinonímia e de tipos de dados, perfazendo 151 heterogeneidades, o que representa 73% das heterogeneidades da aplicação. As heterogeneidades de hiperonímia, que totalizam 36 casos, e as heterogeneidades de atributos compostos, que totalizam 20 casos, não foram detectadas pela COMA 3.0 CE, pois a ferramenta não tem algoritmos para detectar essas heterogeneidades.

Tabela 2. Exemplos de heterogeneidades do estudo de caso

Nome dos campos do esquema 1	Nome dos campos do esquema 2	Heterogeneidade	Descrição
cd_tributo	COD_ATVD_COML	Sinonímia	Código de tributo é sinônimo de código de atividade comercial
cd_contribuinte_mobiliario	COD_INSC_MNUL	Sinonímia	Código do contribuinte mobiliário é sinônimo de código de inscrição municipal
NOM_ENDE_SQL	tp_logradouro, nm_logradouro	Atributo composto	Endereço é desdobrado em tipo e nome de logradouro
nr_cgc_cpf_contribuinte	COD_CPF_RAIZ_CNPJ, COD_ CPLM_ CNPJ	Atributo composto	CNPJ é desdobrado em CNPJ raiz e complemento de CNPJ
UnidAgente	cd_subinspetoria	Hiperonímia	Unidade do agente é hiperônimo de subinspetoria
nm_contribuinte	nm_proprietario	Hiperonímia	Contribuinte é hiperônimo de proprietário

4.3. Técnica usada para alinhar heterogeneidades não detectadas pela COMA 3.0 CE

Para alinhar as heterogeneidades de atributo composto e de hiperonímia, aplicou-se uma técnica elementar externa que utiliza uma base de conhecimento. Conforme [Euzenat e Shvaiko 2010], essa técnica pode ser aplicada da forma como propõe [Aleksovski et al. 2006], que usa como referência uma base de conhecimento de um domínio específico, estruturada por meio de uma ontologia. Conforme [Aleksovski et al. 2006], a obtenção do alinhamento é subdividida em duas etapas: ancoragem e derivação de relações. Inicialmente, na etapa de ancoragem, cada esquema em comparação é alinhado a uma base de conhecimento de um domínio específico. Os elementos da base de conhecimento alinhados aos elementos dos esquemas são denominados âncoras. Na etapa seguinte, de derivação de relações, verifica-se qual o relacionamento existente entre as âncoras e, com base nessa relação, induz-se qual é a relação entre os elementos dos esquemas que foram alinhados às âncoras.

Essa técnica foi aplicada no estudo de caso de forma semelhante ao proposto por [Aleksovski et al. 2006]. Inicialmente, é feita uma busca, em uma ontologia do domínio da aplicação, por elementos com propriedades específicas de atributos compostos ou hiperônimos. Encontradas essas propriedades, verifica-se se os valores (ou instâncias) das propriedades coincidem com nomes dos campos dos esquemas em comparação. Os valores das propriedades são possíveis nomes dos campos desses esquemas. Caso os nomes dos campos dos esquemas comparados coincidam com valores das propriedades, esses valores são as âncoras e alinham-se os campos dos esquemas conforme a propriedade que inter-relaciona as âncoras. Desta forma, a busca por âncoras é feita apenas nas instâncias das propriedades específicas, evitando-se a busca em toda a base de conhecimento. Os passos para aplicação da técnica no estudo de caso são representados por meio do diagrama de atividades da figura 1.

Representou-se a base de conhecimento usando o *Resource Description Framework* (RDF) [Manola e Miller 2004]. Para cada heterogeneidade não identificada pela COMA 3.0 CE, as tabelas e campos correspondentes foram mapeados a classes e propriedades, respectivamente. Os campos com heterogeneidade de atributo composto foram mapeados a propriedades compostas, representadas em RDF por meio de nós em branco [Manola e Miller 2004]. Os campos com nomes hiperônimos foram mapeados a propriedades, representadas em RDF por elementos relacionados entre si pela propriedade **rdfs: subPropertyOf**. As representações em RDF foram validadas por

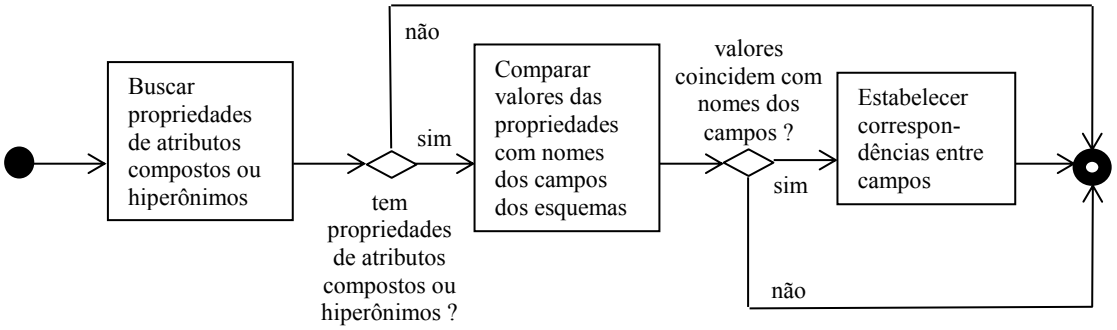

Figura 1. Diagrama de atividades para a técnica aplicada no estudo de caso.

auditores fiscais da Secretaria de Finanças da Prefeitura de São Paulo, com formação na área de computação.

Na figura 2, representam-se algumas classes e propriedades resultantes desse mapeamento. As tabelas da aplicação correspondentes a cadastro de contribuintes, pagamentos de tributos e operações fiscais, por exemplo, foram mapeadas a classes, representadas na figura 2 pelos nós "Contribuinte", "Pagamento de tributo" e "Operação fiscal", respectivamente. Na figura 2, representa-se a propriedade composta "data de pagamento" por meio do nó em branco A0, que se desdobra nas propriedades "dia de pagamento", "mês de pagamento" e "ano de pagamento", representadas por setas. Os nós "unidade do expediente" e "subinspetoria" representam propriedades e "unidade do expediente" é um hiperônimo de "subinspetoria". Nomes de campos dos esquemas foram mapeados a essas propriedades e estão representados na figura 2 por retângulos.

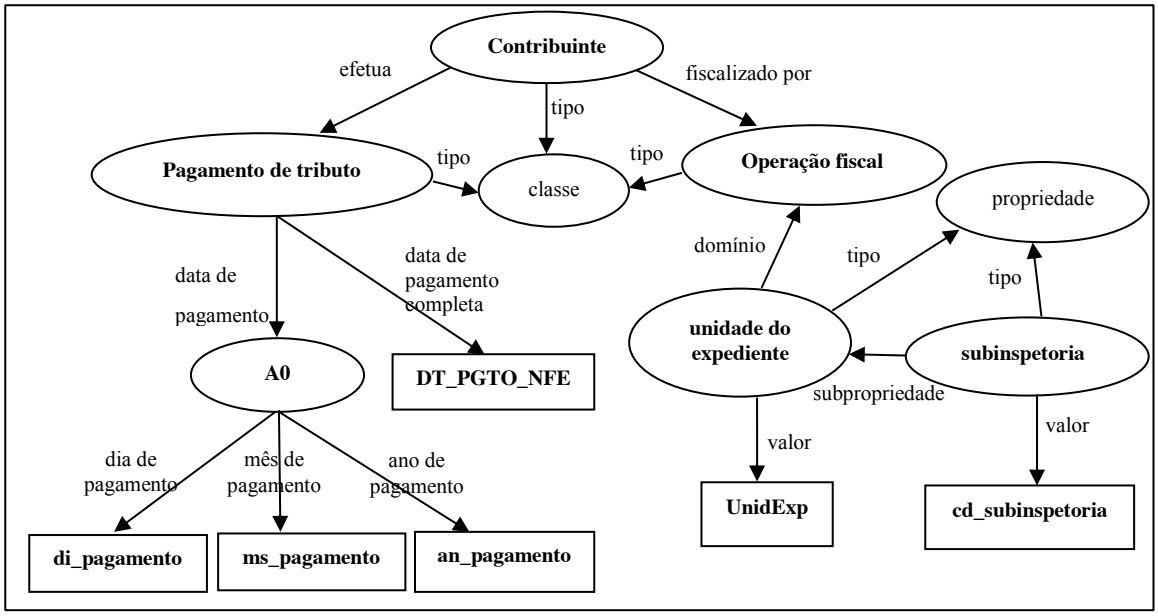

Figura 2. Exemplo de representação de classes e propriedades resultantes do mapeamento a tabelas e campos com heterogeneidades.

Foi implementado um algoritmo para atributos compostos e outro para hiperônimos, em um sistema baseado no padrão de projeto *Model-View-Controller* [Reenskaug 2003]. Na implementação, foi usada a linguagem de programação Java [Java 2012]. Para efetuar a leitura de arquivos em RDF, usou-se a biblioteca Jena [Jena 2012]. Os algoritmos foram posteriormente integrados à COMA 3.0 CE.

4.4. Resultados com a utilização dos algoritmos propostos após sua integração à COMA 3.0 CE

Verificou-se que a aplicação dos algoritmos propostos permitiu detectar as heterogeneidades não alinhadas pela COMA 3.0 CE no estudo de caso (atributos compostos e hiperônimos). Ao integrar os algoritmos propostos à ferramenta, todas as heterogeneidades da aplicação, detectadas visualmente, foram alinhadas pela COMA 3.0 CE. Ou seja, o percentual de heterogeneidades identificadas passou de 73%, sem o uso dos algoritmos propostos, para 100%, ao serem usados também os algoritmos propostos.

5. Conclusões e trabalhos futuros

A obtenção do alinhamento é o passo inicial no processo de integração de bancos de dados e o alinhamento é dado de entrada para ferramentas de reescrita de consultas, passo seguinte no processo de integração. Portanto, o alinhamento é fundamental para o processo de integração. Assim sendo, o alinhamento obtido no estudo de caso contribui para o processo de integração dos bancos de dados tributários da Prefeitura de São Paulo. Essa integração, por sua vez, é importante tanto para os contribuintes como para a administração tributária, pois permite o desenvolvimento de um sistema tributário integrado. Os sistemas tributários atualmente existentes na Prefeitura de São Paulo estão em fase de testes de integração.

Da análise dos resultados obtidos, conclui-se que o uso de uma técnica baseada em uma ontologia do domínio contribuiu para obter o alinhamento de heterogeneidades de atributos compostos e de hiperonímia, que correspondem a 27% das heterogeneidades do estudo de caso. A COMA 3.0 CE não tem algoritmos para essas heterogeneidades e a integração dos algoritmos propostos à ferramenta contribuiu para a extensão da COMA 3.0 CE.

Como trabalhos futuros, sugere-se aplicar uma ferramenta de reescrita de consultas ao alinhamento obtido no estudo de caso. As consultas geradas pela ferramenta de reescrita são aplicadas aos bancos de dados do estudo de caso e técnicos tributários analisam os resultados das consultas, comparando-os com dados obtidos dos sistemas tributários.

Referências

Aleksovski, Z.; Klein, M.; Kate, W. T.; Harmelen, F. V. (2006) “Matching Unstructured Vocabularies using a Background Ontology”, In: Proc.15 th International Conference on Knowledge Engineering and Knowledge Management (EKAW), 2006.

Castano, S.; Antonellis, V.; Vimercati, S.C (2001) “Global Viewing of Heterogeneous Data Sources”, IEEE Transactions on Knowledge and Data Engineering, vol. 13, n.º 2, p. 277-297.

Coma 3.0 Community Edition (2012). Página web. Disponível em http://sourceforge.net/p/coma-ce/wiki/Home/. Acesso em 11/06/2012. Universidade de Leipzig, Alemanha.

Ehrig, M. e Staab, S. (2004) “QOM: Quick Ontology Mapping”, In: Proc. International Semantic Web Conference (ISWC), p. 683-697.

Ehrig, M. e Sure, Y. (2004) “Ontology Mapping – an integrated approach”, In: Proc. European Semantic Web Symposium (ESWS), p. 76-91.

Euzenat, J. e Shvaiko, P. (2010) “Ontology Matching”, Springer.

Euzenat, J. e Valtchev, P. (2004) “Similarity-based ontology alignment in OWL-Lite”, In: Proc. European Conference on Artificial Intelligence (ECAI), p. 333-337.

Hai, D. H. (2005) “Schema Matching and Mapping-based Data Integration”. Tese. Universidade de Leipzig, Alemanha, 222 p..

Hamming, R. (1950) “Error detecting and error correcting codes”, In: Technical Report 2, Bell System Technical Journal.

Java (2012). Página web. Disponível em http://www.java.com/pt_BR/. Acesso em 02/08/2012. Oracle.

Jena (2012). Página web. Disponível em http://jena.apache.org/. Acesso em 02/08/2012. The Apache Software Foundation.

Kim, W.; Seo, J. (1991) “Classifying Schematic and Data Heterogeneity in Multidatabase Systems”, Computer, Los Alamitos, vol. 24, p. 12-18.

Madhavan, J.; Bernstein, P.; Rahm E. (2001) “Generic Schema Matching with Cupid”, In: Proc. 27th International Conference on Very Large Data Bases (VLDB), p. 48-58.

Manola, F. e Miller, E. (2004) “RDF Primer W3C Recommendation”. Disponível em http://www.w3.org/TR/2004/REC-rdf-primer-20040210/. Acesso em 17/12/2011.

Massmann, S.; Raunich, S; Aumüller, D.; Arnold, P.; Rahm, E. (2011) “Evolution of the COMA Match System”, In: The Sixth International Workshop on Ontology Matching.

Miller, G. A. (1993) “Nouns in Wordnet: A Lexical Inheritance System”. Disponível em http://wordnetcode.princeton.edu/5papers.pdf. Acesso em 20/05/2012.

Noy, N. F.; Musen, M. A. (2001) “Anchor-PROMPT: Using Non-Local Context for Semantic Matching”, In: Proc. of the International Joint Conference on Artificial Intelligence (IJCAI), p. 63-70.

Reenskaug, T. M. H. (2003) “The Model-View-Controller. Its Past and Present”. Disponível em http://heim.ifi.uio.no/~trygver/2003/javazone-jaoo/MVC_pattern.pdf. Acesso em 31/07/2012.

Schema and Ontology Matching with Coma 3.0 (2012). Página web. Disponível em http://dbs.uni-leipzig.de/en/Research/coma.html/. Acesso em 04/03/2013. Universidade de Leipzig, Alemanha.

Shvaiko, P.; Giunchiglia, F.; Yatskevich, M. (2009) “Semantic Matching with S-Match”, Technical Report.

O trabalho descrito neste artigo foi desenvolvido no Instituto de Pesquisas Tecnológicas do Estado de São Paulo.