# Interoperabilidade de Operações de Limpeza de Dados Recorrendo a Ontologias



Ricardo Almeida, Paulo Oliveira

Departamento de Engenharia Informática, Instituto Superior de Engenharia – Instituto Politécnico do Porto

GECAD – Grupo de Investigação em Engenharia do Conhecimento e Apoio à Decisão

Porto, Portugal

ral@isep.ipp.pt, pjo@isep.ipp.pt

*Resumo*-**O surgimento de novos modelos de negócio, nomeadamente o estabelecimento de parcerias entre organizações, a possibilidade de as empresas poderem adicionar informação existente na web, em especial na web semânjtica, à informação de que dispõem, levou ao acentuar de alguns problemas já existentes nas bases de dados, nomeadamente no que respeita a problemas de qualidade de dados.**

**Dados de má qualidade podem levar à perda de competitividade das organizações que os detêm, podendo inclusive levar ao seu desaparecimento, uma vez que muitas das suas tomadas de decisão são baseadas nestes dados. Por este motivo torna-se relevante a limpeza de dados que é um processo automático de detecção e eventual correção dos problemas de qualidade de dados, tais como registos duplicados, dados incompletos, e/ou inconsistentes. As abordagens atualmente existentes para solucionar estes problemas, encontram-se muito ligadas ao esquema das bases de dados e a domínios específicos. Para que as operações de limpeza de dados possam ser utilizadas em diferentes repositórios, torna-se necessário o entendimento, por parte dos sistemas computacionais, desses mesmos dados, ou seja, é necessária uma semântica associada. A solução apresentada passa pelo uso de ontologias, como forma de representação das operações de limpeza, para solucionar os problemas de heterogeneidade semântica, quer ao nível dos dados existentes nos diversos repositórios, quer ao nível da especificação das operações de limpeza. Estando as operações de limpeza definidas ao nível conceptual e existindo mapeamentos entre as ontologias de domínio e a ontologia associada a uma qualquer base de dados, aquelas poderão ser instanciadas e propostas ao utilizador para serem executadas sobre essa base de dados, permitindo assim a sua interoperabilidade.**

***Palavras chave: Interoperabilidade, Ontologias, Qualidade de dados, Limpeza de Dados***

## I. INTRODUÇÃO

Com a globalização da economia Mundial, surgiram uma série de alterações ao nível das empresas, nomeadamente, no que respeita aos modelos de negócio, através da fusão, ou do estabelecimento de parcerias entre diferentes empresas. Estas operações, sejam elas, parcerias ou fusões, podem ocorrer entre empresas do mesmo ramo, ou de ramos diferentes e pertencentes ao mesmo país, ou a diferentes países de origem. Estas alterações vão-se refletir, também, ao nível das bases de dados das respectivas empresas, pois, cada uma das empresas possui a sua base de dados, com o seu esquema e dados associados. Daqui resulta, normalmente, uma heterogeneidade em termos de tipos de dados, formatos, bem como heterogeneidade semântica, ou seja, dados com diferentes significados em diferentes domínios (ex: vela no domínio dos automóveis e vela no domínio dos barcos). Torna-se assim necessário, não só a integração das suas bases de dados individuais, bem como das suas regras de negócio de forma a garantir a qualidade dos mesmos.

Problemas de qualidade de dados, tais como:, valores em falta em atributos supostamente obrigatórios; violações de domínio; violações à unicidade dos valores do atributo; violações de regras de negócio; existência de registos duplicados; registos inconsistentes [Oliveira, 2008]; existem em repositórios individuais, bem como na integração de repositórios de dados. Estes problemas assumem uma maior dimensão na integração de repositórios, uma vez que:

- Os mesmos dados podem existir em diferentes repositórios/bases de dados;
- Os dados existentes nos diferentes repositórios encontram-se representados de forma diferente;
- Há uma grande quantidade de informação a ser processada.

A limpeza de dados significa detecção e correção automática dos problemas de qualidade existentes nos dados. Tornar este processo num processo totalmente automático , isto é, sem envolver o ser humano é uma tarefa muito complexa. Como exemplo, podemos imaginar uma situação em que um determinado paciente de um Hospital tem dois registos com moradas diferentes. O que fazer neste caso? Manter os dois registos? Ou eliminar?

A especificação de quais os problemas de qualidade de dados cuja existência se pretende verificar, a forma de o fazer, e as ações corretivas a realizar dependem de conhecimento especializado/pericial. No entanto, uma vez especificado o conhecimento de limpeza de dados, deve tirar-se partido deste, sugerindo ao utilizador a execução de uma dada operação de limpeza sempre que seja possível reconhecer que se está perante mais uma situação semelhante. Desta forma, ao ser humano compete apenas aceitar ou rejeitar a execução das operações de limpeza de dados que lhe são sugeridas. Pelo que se conhece, não há qualquer protótipo de investigação ou ferramenta comercial que armazene conhecimento de limpeza

de dados sobre o domínio e se socorra deste para propor a execução de operações de limpeza.

As ferramentas existentes de limpeza de dados, sejam estas académicas (e.g., Ajax [Galhardas et al., 2000]; Arktos II [Vassiliadis et al., 2003]; IntelliClean [Low et al., 2001]; Potter's Wheel [Raman and Hellerstein, 2001]; FraQL [Sattler e Schallehn, 2001]; e, SmartClean [Oliveira, 2008]) ou comerciais (e.g., Trillium Quality [Trillium, 2011]; e, DataFlux [DataFlux, 2011]), dependem totalmente do utilizador para a especificação das operações de detecção e correção a efetuar. O utilizador começa por definir as operações de detecção a executar, na procura dos eventuais problemas de qualidade que possam existir nos dados. Após a execução destas, o utilizador especifica as operações de correção a realizar nos dados para solucionar os problemas identificados. Tipicamente, a realização de um processo de limpeza de dados envolve a especificação manual de um número elevado de operações de detecção e correção.

Embora não seja possível conceber uma solução mágica para a limpeza de dados [Raman and Hellerstein, 2001], isto é, um sistema completamente automático que detecte e corrija os problemas de qualidade existentes nos dados, sem necessitar de qualquer intervenção humana, esta pode ser substancialmente reduzida comparativamente ao que atualmente acontece.

A abordagem seguida, até ao momento, nos protótipos de investigação e nas ferramentas comerciais, consiste na especificação das operações de limpeza de dados ao nível do esquema dos dados. As operações são especificadas com base nos nomes dos atributos, das tabelas e das bases de dados. A abordagem é adequada caso a limpeza de dados seja executada apenas numa única base de dados cujo esquema permaneça inalterado.. No entanto, na maioria das vezes isto não corresponde à realidade. A generalidade das operações de limpeza de dados são genéricas o suficiente para poderem ser executadas em bases de dados diferentes num mesmo domínio. No limite, uma operação de limpeza de dados pode ser de tal forma genérica que até é independente de um domínio específico (e.g., detecção de violação de sintaxe num atributo que armazena códigos postais Portugueses). Nestes casos, a abordagem que tem sido seguida não é adequada, uma vez que "prende" as operações ao esquema de uma base de dados específica. Atendendo a que as operações foram especificadas para uma base de dados em concreto, não é trivial efetuar a sua execução noutra base de dados, uma vez que o esquema de dados não é o mesmo. A execução das operações está pois, condicionada à realização de um conjunto de alterações.

Assim, pretende-se minimizar a intervenção humana no processo de especificação das operações de limpeza de dados, de modo a que este não seja tão manual e dependente do ser humano, recorrendo a duas vias: (i) reutilizar operações de limpeza de dados entre bases de dados diferentes; e, (ii) tirar partido do conhecimento de limpeza de dados já existente sobre os domínios, para propor ao utilizador um conjunto de operações cuja execução seja relevante para a detecção e correção dos problemas de qualidade existentes na sua base de dados.

Este artigo encontra-se organizado da seguinte forma. No segundo capítulo são apresentados trabalhos relacionados com a qualidade de dados. No terceiro capítulo são apresentados os problemas de qualidade a serem trabalhados. No quarto capítulo é apresentada uma proposta de modelo que visa a interoperabilidade das operações de limpeza e efectuada a sua descrição. Finalmente, no capítulo cinco são apresentadas as conclusões e trabalho futuro.

## II. TRABALHO RELACIONADO

Existem alguns trabalhos nesta área, no que respeita à utilização de ontologias para a Gestão de Qualidade de dados (*Data Quality Management*) e que podem ser divididas em duas grandes áreas:

- Uso de ontologias de domínio:
  - Kedad e Métais (2002) referem que em sistemas de informação "multi-fonte", nomeadamente em armazéns de dados, existem dois tipos de heterogeneidade: a *intensional* (intensional) e a *extensional* (extensional – data cleaning). A heterogeneidade *intensional* prende-se com conflitos relacionados com a estrutura dos dados, enquanto, na *extensional*, estes conflitos se situam ao nível das instâncias de dados. Eles apresentam uma solução de limpeza de dados ao nível das instâncias no conhecimento linguístico fornecido por uma ontologia de domínio.
  - Milano, Scannapieco, e Catarci, (2005). Neste caso, as ontologias são utilizadas com o intuito de fornecerem o conhecimento específico acerca de um determinado domínio de forma a permitir a validação e limpeza dos dados ao nível do conhecimento acerca de um determinado domínio.
  - Rey, Anguita e Crespo (2006) desenvolveram também um sistema de descoberta de conhecimento (OntoDataClean) baseado em ontologias e que permite o pré-processamento de dados e a integração ao nível das instâncias. Segundo eles, dependendo de se tratar de uma única, ou várias fontes de dados, as inconsistências podem ser, respectivamente, ao nível da instância ou ao nível do esquema.
  - Brueggemann e Gruening (2008) sugerem a utilização do conhecimento fornecido pelas ontologias no contexto da Gestão de Qualidade de Dados, nomeadamente no que diz respeito a problemas de: consistência, detecção de duplicados e gestão de metadados. Esta abordagem estende a efectuada por Milano, Scannapieco, e Catarci, (2005), pois é efectuada a anotação das ontologias de domínio com metadados;
  - Zhang e Wang (2008), também sugerem o uso de ontologias para o ETL (Extraction, Transformation and Loading), pois estas podem ser partilhadas, reutilizadas e estruturadas semanticamente.
  - Cai e Xu (2010) sugerem o uso de ontologias de domínio no processo de ETL para encontrar as fontes de dados, definindo as regras de

transformação dos dados, e eliminando possíveis heterogeneidades. Nesta abordagem, a ontologia do domínio encontra-se embebida nos metadados do armazém de dados.

  - Furber e Hepp (2011) fornecem um modelo conceptual que permite a representação de regras de qualidade usando RDF e OWL, o que permite a reutilização das mesmas em problemas de qualidade de dados.

- Gestão de métodos e problemas da qualidade de dados:
  - Nesta área podemos encontrar o OntoClean (Wang, Hamilton, Bither, e Science, 2005), que fornece um template para a limpeza de dados e se encontra dividido em vários passos tais como a construção de uma ontologia, tradução dos objectivos do utilizador para a limpeza de dados em linguagens de consulta às ontologias e seleção de algoritmos de limpeza.
  - Fürber e Hepp (2010) representam o conhecimento em ontologias anotadas diretamente com metadados específicos da gestão de qualidade de dados. A abordagem efectuada por estes autores visa a utilização da estrutura das ontologias de domínio para o fornecimento de sugestões de dados inválidos, identificação de duplicados e para a anotação da qualidade de dados quer ao nível do esquema, quer ao nível das instâncias.

Após a análise efectuada relativamente a trabalhos existentes na área, constata-se que os mesmos se encontram muito virados para problemas de qualidade de dados em domínios específicos, o que limita a sua aplicação, pois existem vários problemas de dados que são genéricos e não podem ser utilizados. Uma outra lacuna identificada prende-se com a possibilidade da utilização de mapeamentos existentes entre ontologias para a reutilização de operações existentes. Sempre que um perito estabelecer um mapeamento entre a ontologia associada à sua base de dados e a ontologia de domínio, devem ser-lhe sugeridas, não só, as operações existentes nesta, mas também, outras com as quais estes conceitos se encontrem associados através de mapeamentos com outras ontologias. Na secção seguinte irão ser abordados quais os problemas de qualidade de dados existentes nas instâncias e que serão objeto de estudo no âmbito do trabalho que se encontra em curso.

## III. PROBLEMAS DE QUALIDADE DE DADOS

As operações de limpeza de dados destinam-se a problemas de qualidade existentes nas instâncias [Oliveira, 2008], nomeadamente:

- Ao nível do atributo:
  - Valor individual:
    - Valor em falta em atributos de preenchimento obrigatório;
    - Violação de sintaxe. Qual a sintaxe esperada para o atributo. Um número de contribuinte tem que ser composto por nove dígitos numéricos;
    - Violação de domínio: Pode ocorrer num intervalo de valores ou; Não respeitar um conjunto de valores possíveis;
  - Múltiplos valores do atributo:
    - Sinónimos;
    - Violação de unicidade. Esta restrição prende-se com a cardinalidade 1;
    - Violação de restrição de integridade relativamente a um determinado atributo.
- Ao nível do tuplo, podendo os dados representados no tuplo não serem consistentes (ex: Total=qtdade*valor);
- Ao nível da relação:
  - Violação de dependência Funcional. Para tabelas que não se encontrem na 3ª forma normal (ex: existirem vários registos onde o código postal e localidade são 4000 Porto e existir um registo onde o código postal é 4000 e localidade Lisboa);
  - Circularidade entre tuplos – auto relacionamento (ex: A é patrão de B e B é patrão de A);
  - Tuplos duplicados
  - Violação de restrição de integridade (ex: um número de factura com uma data superior ser inferior a um número de factura com uma data anterior).
- Múltiplas relações ao nível das instâncias:
  - Heterogeneidade de sintaxes (ex: data no formato Português e data no formato Americano);
  - Heterogeneidade de unidades de medida;
  - Sinónimos;
  - Homónimos (ex: vela de barco vs vela de automóvel);
  - Diferentes granularidades de representação (ex: numa base de dados a tabela estado civil é composta por: solteiro, casado, viúvo e união de facto e noutra base de dados a tabela estado civil é representada por: solteiro, casado);
  - Violação de integridade referencial;
  - Tuplos duplicados;
  - Violação de restrição de integridade. (ex: uma empresa que tem delegações em Lisboa e no Porto. Existe um funcionário que tem 2 projetos em cada delegação, sendo que é regra da empresa que nenhum funcionário pode estar em mais de 3 projetos ao mesmo tempo).

## IV. SOLUÇÃO PROPOSTA

A solução proposta visa permitir a reutilização de operações de limpeza respeitantes aos problemas de qualidade de dados apresentados na secção anterior. Atualmente, as operações encontram-se associadas ao esquema das bases de dados. Pretende-se que estas operações possam ser utilizadas em novas bases de dados, mas minimizando a intervenção humana no processo. Para que se consiga a reutilização torna-se necessário que a especificação das operações de limpeza seja feita numa linguagem formal e a um nível conceptual de forma a permitir a sua "independência" relativamente a um esquema em particular, o que irá permitir a sua interoperabilidade. Desta forma sugere-se a utilização de ontologias, utilizando o OWL (2011) para a sua representação e especificação das operações de limpeza, sempre que a sua expressividade o permita. Nas restantes situações utiliza-se o SWRL (2011). O SWRL é uma linguagem de regras para a web semântica, baseada em OWL-DL e OWL lite. As regras são definidas através de conceitos OWL (classes, propriedades, indivíduos, entre outros). As regras criadas podem ser depois armazenadas nas próprias ontologias. As regras desta linguagem não suportam a negação, nem disjunção de átomos.

De forma a melhor ilustrar a abordagem que se propõe recorre-se a um exemplo, simplificado, de uma ontologia de domínio na área do ensino superior, representada em UML. Das várias classes, aquela que contem maior número de propriedades é a classe Pessoa, em virtude das regras de limpeza apresentadas posteriormente incidirem sobre esta.

Relativamente às operações de limpeza, estas podem ser genéricas ou específicas e independentes ou dependentes do domínio. No caso das operações de limpeza serem

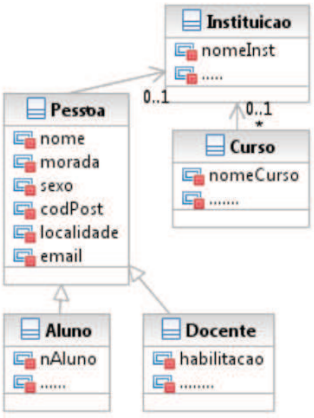


Figura 1 Representação UML da ontologia exemplo

independentes do domínio, tais como, por exemplo, a detecção de valores não permitidos (isto é, que violam o domínio) na propriedade sexo da classe Pessoa. A regra que define esta operação em SWRL pode ser representada da seguinte forma: xsd:string (?sexo); ["masculino", "feminino"] (?sexo). Esta regra genérica pode ser aplicada em qualquer domínio. Numa base de dados em concreto, esta regra necessita de ser materializada aos valores que definem masculino e feminino (ex: m;f).

Caso as operações sejam relativas a um domínio específico, estas, devem ser guardadas nesse mesmo domínio. Como exemplo desta situação, temos número de aluno, que se encontra representado na propriedade nAluno da classe Pessoa. Neste contexto, esta propriedade é de preenchimento obrigatório. Uma vez que o OWL DL possui expressividade suficiente para a representação desta regra, apresenta-se de seguida o extrato correspondente:

```
<owl:Class rdf:ID="Pessoa">

  <owl:Restriction>
    <owl:onProperty rdf:resource="#nAluno"/>
    <owl:minCardinality
rdf:datatype="&xsd;nonNegativeInteger">1</owl:minCardinal
ity>
   </owl:Restriction>
….
</owl:Class>
```

Figura 2 Arquitetura do modelo proposto

Para que esta solução possa ser aplicada a diferentes bases de dados é necessária a conversão do esquema da base de dados, sobre o qual se pretendem efectuar as operações de limpeza, numa ontologia, utilizando uma ferramenta como o Protégé.

As operações de limpeza representadas desta forma podem agora ser efectuadas em diferentes bases de dados, bastando para isso a sua "transformação" para uma ferramenta de limpeza de dados (ex: SmartClean [Oliveira, 2009]).

Assim, na sequência do atrás exemplificado, propõe-se o modelo apresentado na Figura  e que será agora descrito. O modelo é composto por uma ontologia ortogonal de limpeza (OLo), onde são definidas as classes, propriedades e regras a que estas devem obedecer, que sejam genéricas e independentes do domínio. Um perito / especialista (pessoa responsável pela limpeza de dados), estabelece mapeamentos entre a ontologia ortogonal de limpeza e a(s) ontologia(s) de domínio (OLd(n)), ao nível de conceitos e propriedades

comuns, o que permitirá a aplicação da regra (genérica e independente) a uma determinada propriedade de um conceito de um domínio (específico). Desta forma, sempre que um perito estabeleça um mapeamento entre uma ontologia resultante de uma base de dados e a ontologia ortogonal de limpeza e/ ou ontologia(s) de domínio, tendo por base as regras existentes, é possível propor/sugerir ao perito, as operações de limpeza de dados a realizar.

O perito/especialista pode também estabelecer mapeamentos diretos entre a ontologia associada à base de dados alvo de limpeza, e uma outra ontologia já existente de um outro domínio (representados na Figura  sob a forma de Mx e My). Nesta situação, identificou-se que uma dada propriedade de um conceito específico de um determinado domínio também existe no domínio em questão e, como tal, as mesmas regras também devem ser aplicadas. Por exemplo, existindo uma ontologia no domínio do comércio electrónico e estando representada a propriedade e-mail no conceito pessoa, esta terá uma regra associada de obrigatoriedade de preenchimento. No caso da ontologia apresentada na Figura , sendo estabelecido um mapeamento entre a propriedade e-mail dos conceitos aluno ou docente e a propriedade e-mail do conceito pessoa, isso permite que a mesma regra de preenchimento obrigatório seja sugerida ao perito/especialista para posterior execução.

Se as operações sugeridas, ao utilizador, não forem as suficientes para que sejam efectuadas as operações de limpeza pretendidas, estas devem ser definidas por este e posteriormente armazenadas na ontologia de limpeza, caso estas sejam genéricas e supra-domínio. Caso contrário, estas devem ser associadas à ontologia de domínio (OLd(n)) correspondente. Este modelo não é, portanto, estático, nem respeitante a um determinado domínio, mas sim, evolutivo, uma vez que pode evoluir, quer em número de ontologias de domínio, quer ao nível das próprias ontologias, quer em número de operações de limpeza, quer em termos de mapeamentos associados.

No caso das operações de limpeza corresponderem a um determinado domínio (específicas ou dependentes do domínio), estas terão de ser associadas aos conceitos/propriedades definidos nas ontologias de domínio. Caso contrário (genéricas ou independentes do domínio), estas poderão ser armazenadas na própria ontologia ortogonal de limpeza.

As operações de limpeza de dados que resultam dos mapeamentos estabelecidos e que são aceites pelo perito/especialista para execução, passam por um processo de transformação que as coloca ao nível do esquema da base de dados e respeitando a sintaxe utilizada pela ferramenta de limpeza. Compete a este proceder à sua execução.

## V. CONCLUSÕES E TRABALHO FUTURO

Neste artigo, apresentamos um modelo, composto por uma ontologia ortogonal de limpeza (OLo) e ontologias de domínio (OLdn) que visam permitir a representação de operações de limpeza a um nível conceptual de forma a permitir a sua reutilização em diferentes bases de dados. Assim, as operações genéricas e independentes do domínio são definidas na ontologia OLo e as dependentes do domínio em OLd e as operações nelas existentes podem ser reutilizadas. O estabelecimento de mapeamentos permite uma equivalência semântica entre conceitos e propriedades das diferentes ontologias, o que suporta que operações de limpeza anteriormente definidas possam ser propostas/sugeridas ao perito/especialista para execução. O trabalho a realizar a partir deste momento passa por uma especificação mais detalhada das ontologias e mapeamentos. Posteriormente, será efetuada a implementação da solução proposta e a consequente aplicação em casos de estudo.

## REFERÊNCIAS


[1] Galhardas, H., Florescu, D., Shasha, D. and Simon, E. (2000) AJAX: An Extensible Data Cleaning Tool. In Proceedings of the ACM SIGMOD on Management of Data, Dallas (USA), May of 2000. pp. 590.

[2] DataFlux, "Data Flux Products". Available at http://www.dataflux.com/Products /Products.aspx#studio, in February 14th of 2011 at 12:25.

[3] Oliveira, Paulo (2008) - Detection and Correction of Data Quality Problems: Model, Syntax and Semantic. School of Engineering of University of Minho, PhD Thesis in Computer Science, Sept 2008.

[4] Raman, V. and Hellerstein, J. M. (2001). Potter's Wheel: An Interactive Data Cleaning System. In Proceedings of the 27th Very Large Databases Conference, Roma (Itália), September of 2001. pp. 381-390.

[5] Trillium Software, "TS Quality". Available at http://www.trilliumsoftware.com/home/products/TSQuality.aspx, in February 14th of 2011, at 10:15.

[6] Vassiliadis, P.,Simitsis, A., Georgantas, P. and Terrovitis, M. (2003). A Framework for the Design of ETL Scenarios. In Proceedings of the 15th Conference on Advanced Information Systems Engineering (CAiSE'03), Klagenfurt (Austria), June of 2003. pp. 520-535.

[7] SWRL. Available at http://www.w3.org/Submission/SWRL/, in February 16 th of 2011, at 10:00.

[8] Oliveira, P. ; Rodrigues, F. e Henriques, P. – "SmartClean: An Incremental Data Cleaning Tool". In Proceedings of the 9th Int. Conference on Quality Software, p. 452-457. Jeju (Korea), August 2009.

[9] OWL. Available at http://www.w3.org/TR/owl-guide/, in February 16 th of 2011, at 15:32.

[10] Brueggemann, S., and Gruening, F. (2008). Using Domain Knowledge Provided by Ontologies for Improving Data Quality Management. Studies in Computational Intelligence, 2009, Volume 221/2009, 187-203

[11] Fürber, C., and Hepp, M. (2010). Using semantic web resources for data quality management. Knowledge Engineering and Management by the Masses, 211–225.

[12] Fürber, Christian, and Hepp, M. (2011). Towards a vocabulary for data quality management in semantic web architectures. Proceedings of the 1st Int. Workshop on Linked Web Data Management, LWDM '11 (pp. 1–8). New York, NY, USA: ACM. doi:10.1145/1966901.1966903

[13] Jiang, L., Cai, H., and Xu, B. (2010). A Domain Ontology Approach in the ETL Process of Data Warehousing. IEEE International Conference on E-Business Engineering (pp. 30–35).

[14] Kedad, Z., and Métais, E. (2002). Ontology-based data cleaning. Natural Language Processing and Information Systems, 137–149.

[15] Milano, D., Scannapieco, M., and Catarci, T. (2005). Using ontologies for xml data cleaning. On the Move to Meaningful Internet Systems 2005: OTM Workshops (pp. 562–571).

[16] Perez-Rey, D., Anguita, A., and Crespo, J. (2006). Ontodataclean: Ontology-based integration and preprocessing of distributed data. Biological and Medical Data Analysis, 262–272.

[17] Wang, X., Hamilton, H. J., Bither, Y., and Science, U. of R. D. of C. (2005). An ontology-based approach to data cleaning. Citeseer.

[18] Zhuolun Zhang, and Sufen Wang. (2008). A Framework Model Study for Ontology-Driven ETL Processes. Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th Int. Conf. on (pp. 1-4).