Avaliação da Qualidade em Linked Datasets: uma abordagem com foco nos requisitos da aplicação

Aluno: Walter Travassos Sarinho[1]

Orientadora: Bernadette Farias Lóscio[1]

Co-Orientadora: Damires Souza[2]

[1]Programa de Pós Graduação em Ciências da Computação – Centro de Informática (CIn) – Universidade Federal de Pernambuco (UFPE)
Recife – Pernambuco – Brasil

{wts, bfl}@cin.ufpe.br

[2]Instituto Federal de Educação, Ciência e Tecnologia (IFPB)
João Pessoa – Paraíba – Brasil

damires@ifpb.edu.br

Nível: Mestrado

Ano de Ingresso no Programa: 2012

Época Esperada de Conclusão: Fevereiro de 2014

Etapas Concluídas: Referencial Bibliográfico (Dez/2012), Definição do Problema (Mar/2013), Definição da Arquitetura (Abr/2012), Especificação dos componentes (Jun/2012).

Etapas Futuras: Implementação da Arquitetura (Ago/2013), Experimentação (Set/2013), Escrita da Dissertação (Dez/2013), Defesa da Dissertação (Fev/2014)

Abstract. *The increasing availability of datasets on the Web of Data faces some new challenges. One of them regards how to evaluate the quality of these datasets, whose solution may use Information Quality criteria. Most quality criteria are abstract concepts and need to be calibrated within a good situation before being used to evaluate any task. In this context, this work proposes an approach to evaluate the quality of linked datasets, i.e., datasets published according to the principles of Linked Data. Furthermore, we intend to evaluate the quality considering the application´s requirements. One distinguishing issue of our approach is the use of an extensible repository of quality criteria, which can be defined by a domain expert of the application. As a result, a quality score is generated and a ranking of the linked datasets is produced. This classification intends to serve as a comparative metric for the selection of the best datasets to be used in the process of rewriting queries and data integration.*

Keywords*: Information Quality, Linked Datasets, Semantic Web.*

1. Introdução e Motivação

A Web é um vasto repositório de dados estruturados, semi ou não estruturados, que cobrem os mais variados domínios do conhecimento. De maneira mais específica, destacam-se os conjuntos de dados disponíveis em RDF (*Resource Descripton Framework* [W3C, 2013]) e publicados de acordo com os princípios de *linked data* [Bizer *et al*., 2009], chamados de *linked datasets*. O interesse na publicação de *linked datasets* é crescente, uma vez que a natureza estruturada e o uso de vocabulários incrementam o nível semântico do dado facilitando sobremaneira o processamento deles por agentes de software. Por um lado, o imenso crescimento na disponibilidade desses dados tem motivado a ideia de desenvolver aplicações que façam uso de múltiplas *linked datasets* [Lóscio *et al*., 2012], por outro, este grande volume de dados e a falta de informações suficientes sobre as fontes trouxeram à tona um grande desafio: a avaliação da qualidade destes *datasets*.

Neste panorama, este trabalho tem como objetivo propor uma solução para a avaliação da qualidade de *linked datasets* de acordo com os requisitos de uma aplicação que consulta dados em múltiplas fontes de dados. Especificamente, *dada uma aplicação A, um conjunto de fontes linked data* $S=\{S_1, S_2,..., S_n\}$ *e um conjunto de consultas* $Q = \{Q_1,..., Q_m\}$*, que a aplicação deseja responder a partir dos dados disponíveis em S*, a abordagem proposta permite *calcular a qualidade das fontes de dados em S* levando em consideração as *consultas em Q*, bem como os demais *requisitos* considerados relevantes do ponto de vista do usuário ou da aplicação. Para mensurar a qualidade de uma fonte de dados, serão utilizados critérios de qualidade consagrados na literatura, como, por exemplo, disponibilidade, precisão e corretude [Zaveri *et al*., 2012; Wang *e* Strong, 1996]. Como resultado do processo de avaliação será obtida uma classificação ordenada das fontes, com suas respectivas medidas de qualidade, que serão obtidas a partir dos requisitos da aplicação e dos critérios de qualidade configurados.

Uma característica destacada nesta proposta é a utilização de um repositório de critérios de qualidade extensível e adaptável que contém diversos critérios para avaliação de *linked datasets*, passíveis de serem configurados por um especialista do domínio da aplicação. Para isso, considera-se que toda aplicação pertence a um domínio de dados, como, por exemplo, dados bibliográficos e dados governamentais. A proposta de ter um repositório adaptável é justificada pelo fato de existirem diversos domínios do conhecimento onde um determinado critério de qualidade pode ser considerado mais importante pelo especialista naquele domínio do que outro critério. O especialista de domínio deve conhecer a proposta da aplicação e seus requisitos para elencar corretamente quais critérios de qualidade devem ser usados, pois o uso indiscriminado de tais critérios de qualidade pode melhorar ou piorar a classificação da qualidade das fontes.

Como principais diferenciais da abordagem proposta, destacam-se: (i) Tem como foco a aplicação, ou seja, considera os requisitos da aplicação na avaliação da qualidade das fontes de dados; (ii) Faz uso de um repositório de critérios de qualidade extensível e configurável e (iii) É adaptável a diferentes domínios do conhecimento.

O restante deste trabalho está organizado como segue: a Seção 2 apresenta a fundamentação teórica; a Seção 3 descreve-se a caracterização da contribuição. Na Seção 4, alguns trabalhos relacionados são abordados e, por fim, na Seção 5, a avaliação dos resultados e estado atual do trabalho são apresentados.

2. Fundamentação Teórica

Qualidade da Informação (QI) é comumente definida como um conjunto de critérios ou dimensões utilizados para indicar o grau de qualidade geral de uma informação obtida por um sistema [Batista 2008; Wang e Strong 1996]. Na literatura, QI é definida como "adequação ao uso" [Wang e Strong, 1996], o que nos leva a considerar que, a informação é apropriada se atende a um conjunto de requisitos estabelecidos, seja por um usuário ou por um conjunto de normas. Dessa forma, o valor da informação depende da sua utilidade [Batista 2008]. Aspectos de QI incluem um conjunto de critérios, métodos de avaliação desses critérios e, normalmente, uma medição geral do grau da QI. Exemplos de critérios de qualidade são: disponibilidade (*availability*) – que verifica se a informação está disponível e alcançável para uso [Zaveri *et al.*, 2012] e precisão (*accuracy*) – mensura o quanto da informação representa corretamente um fato do mundo real [Zaveri *et al*., 2012]. Um critério de grande importância ao nosso trabalho é a completude do esquema (*schema completeness*). Este critério demanda informações sobre o que a aplicação quer consultar nas fontes de dados para se mensurar o quanto uma fonte de dados é completa para responder tais consultas [Zaveri *et al*. 2012].

Em 1996, Wang e Strong (1996) propuseram um *framework* conceitual onde foram agrupados quinze critérios de qualidade em quatro grupos iniciais que segmentam as características da QI em: Contextual, Intrínseca, Representacional e Acessibilidade. Em 2012, Zaveri *et al*. (2012) agrupou mais dois grupos (Confiança e Dinamicidade do *Dataset*) e novos critérios de qualidade ao que foi proposto inicialmente por Wang – totalizando seis grupos e vinte seis critérios de qualidade. O trabalho de Zaveri *et al*. foi realizado levando em consideração *linked datasets.*

Os dados disponíveis e padronizados nas *linked datasets* cobrem os mais diversos domínios, *e.g.* dados geográficos, publicações (dados bibliográficos) e dados governamentais. Contudo, apesar dessa padronização, existem problemas relacionados à qualidade dos dados na Web de Dados como, por exemplo: valores conflitantes entre conjuntos de dados diferentes [Mendes *et al*., 2012], diversidade dos dados [Flemming, 2010], ruído na informação e dificuldade para acessá-la [Hogan *et al*., 2012]. Tais fatores motivam a especificação de critérios de qualidade específicos para *linked datasets,* bem como o desenvolvimento de métricas adequadas para este contexto.

Outro fator importante a ser considerado na avaliação da qualidade dos dados é o domínio ao qual uma aplicação pode pertencer, o qual influencia diretamente na escolha dos critérios no processo de avaliação da qualidade. Por exemplo, o critério idade (*currency*) pode ser útil no domínio de aplicações financeiras, onde o dado deve ser tão atual quanto possível, no entanto, no domínio de dados bibliográficos ele torna-se irrelevante visto que os títulos das publicações são absolutos de acordo com sua data de inserção na fonte de dados. Também se deve considerar que aplicações distintas podem fazer parte de um mesmo domínio e possuir requisitos de aplicação heterogêneos. Isso também leva a uma possível diferença na classificação da qualidade das fontes de dados para cada uma dessas aplicações.

3. Caracterização da Contribuição

Esta seção descreve a abordagem proposta para avaliação da qualidade de *linked datasets*. A arquitetura da Figura 1 ilustra os principais componentes envolvidos no processo de avaliação proposto, os quais são descritos a seguir:

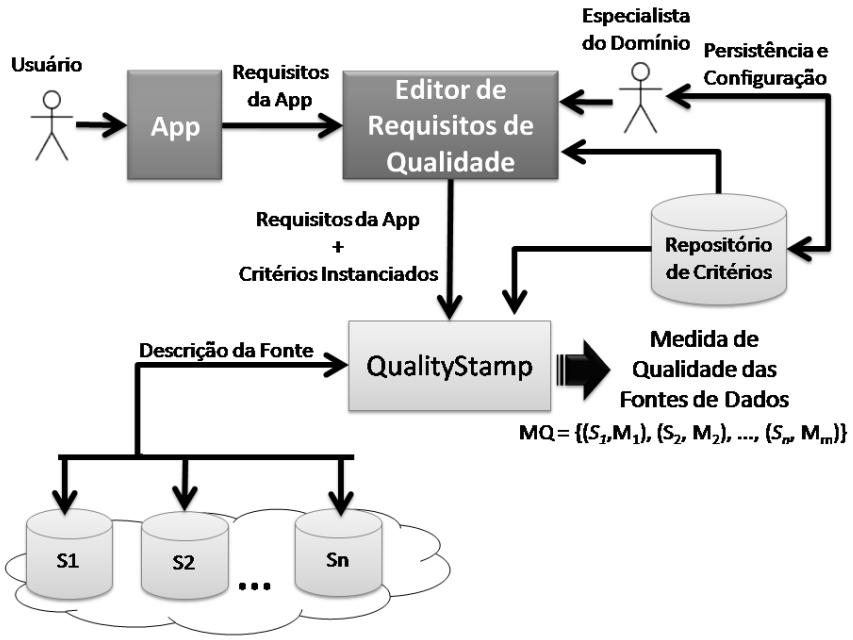

Figura 1 – Arquitetura proposta para avaliação da qualidade das fontes.

Aplicação (App): trata-se de uma aplicação que realiza consultas em um conjunto pré-definido de *linked datasets*. Nesta abordagem, considera-se que as aplicações são de domínio específico, ou seja, as fontes de dados consultadas pela aplicação ($\{S_1, S_2, ..., S_n\}$) pertencem a um mesmo domínio. Além disso, é considerada a presença de um especialista do domínio, o qual será responsável por identificar os critérios de qualidade mais adequados ao domínio em questão. Para cada aplicação, será definido um conjunto de requisitos, os quais podem ser identificados, principalmente, por meio das consultas que a aplicação executa. A partir das consultas, é possível identificar os conceitos do domínio que são relevantes para a aplicação.

Editor de Requisitos de Qualidade: o especialista do domínio, com base nos requisitos da aplicação em questão, usa o *Editor* para verificar quais critérios de qualidade estão disponíveis e configura quais serão instanciados para aquela aplicação. Além disso, é possível atribuir um peso (entre *zero* e *um*) a cada critério de qualidade. Ao escolher o peso *um*, será considerada a totalidade do critério e, optando por *zero*, o critério não será levado em consideração.

Repositório de Critérios de Qualidade: é um repositório central dos critérios de qualidade, o qual armazena os critérios e métodos para cálculo de cada um deles. Ele é extensível, ou seja, capaz de receber novos critérios e métricas. O repositório **disponibiliza** ao *Editor de Requisitos* quais critérios encontram-se disponíveis e, ao *QualityStamp,* como calcular cada critério solicitado pelo *Editor*.

QualityStamp: é o componente principal da arquitetura. Seu nome vem do conhecido selo de qualidade que pode ser atestado a um produto ou serviço. Este componente recebe como parâmetros de entrada as seguintes informações: (i) requisitos da aplicação e critérios instanciados provenientes da *App* e do *Editor de Critérios* respectivamente; (ii) métricas de cálculo obtidas a partir do *Repositório de Critérios* e (iii) Descrições das fontes *linked data* consultadas pela *App*. Como resultado final, o *QualityStamp* gera uma classificação da qualidade das fontes de dados num conjunto de pares (S_n, M_m) tal que: MQ = $\{(S_1,M_1), (S_2, M_2), ..., (S_n, M_m)\}$.

3.1. Visão geral da abordagem proposta

Como mencionado anteriormente, a abordagem proposta tem como objetivo obter uma classificação da qualidade de múltiplos *linked datasets* de acordo com os requisitos de uma aplicação específica. Para isso, inicialmente, o especialista de domínio utiliza o *Editor de Requisitos de Qualidade* para escolher quais critérios são considerados relevantes para o domínio da aplicação. Tais critérios são recuperados a partir do *Repositório de Critérios de Qualidade*. Uma vez que os critérios foram escolhidos, o próximo passo consiste em calcular os valores de cada critério para cada uma das fontes *linked data* $\{S_1,..., S_n\}$. Este cálculo é realizado pelo *QualityStamp* a partir das informações recebidas pelo do *Editor de Requisitos* e das métricas recuperadas a partir do *Repositório de Critérios de Qualidade*. No último passo, o *QualityStamp* calcula a *Medida de Qualidade da Fonte de Dados* para cada fonte *linked data* e, em seguida, apresenta uma classificação geral de qualidade dos *datasets* $\{(S_1,M_1), (S_2, M_2), ..., (S_n, M_m)\}$. A seguir, é apresentado um exemplo com o objetivo de tornar mais clara a proposta.

3.2. Um Exemplo

Como forma de ilustrar a abordagem proposta, será apresentado um exemplo onde se avalia a qualidade de um conjunto de múltiplas fontes *linked data* consultadas por uma aplicação, chamada *Frevo* (***For**m to **Ev**aluate Linked **O**pen Data Sets*). Esta aplicação pertence ao domínio de dados bibliográficos e consulta títulos de artigos a partir do nome de autores, ano de publicação e periódico de publicação do artigo. Especificamente, a aplicação recupera dados submetendo uma mesma consulta SPARQL em três fontes de dados distintas, as quais utilizam a ontologia de referência AKT[1]: IEEE[2], ACM[3] e DBLP[4]. A Tabela 1 apresenta alguns exemplos de consulta que podem ser submetidas pela aplicação.

Tabela 1. Exemplos de consultas da aplicação Frevo

Consulta Q1	Consulta Q2	Consulta Q3

SELECT DISTINCT ?titulo
WHERE {
?artigo akt:has-title ?titulo .
?artigo akt:has-author ?autor .
?autor akt:full-name **"Tim Berners-Lee"**.
}
	SELECT DISTINCT ?titulo
	WHERE {
	?artigo akt:has-title ?titulo .
	?artigo akt:has-author ?autor .
	?autor akt:full-name **"Alon Y. Halevy"**.
	?artigo akt:has-date akt-date:**1993** .
	?artigo akt:article-of-journal ?conferencia .
	?conferencia akt:has-title **"Workshop on Deductive Databases, JICSLP"**}
		SELECT DISTINCT ?titulo
		WHERE {
		?artigo akt:has-title ?titulo .
		?artigo akt:has-author ?autor .
		?autor akt:full-name **"Christian Bizer"**.
		?artigo akt:has-date akt-date:**2003** . }

Considere que para a avaliação da qualidade das fontes IEEE, ACM e DBLP os seguintes critérios, juntamente com seus respectivos pesos, foram escolhidos: (i) disponibilidade (100%); e (ii) completude do esquema (80%). A próxima etapa é o cálculo dos critérios escolhidos. O *QualityStamp* calcula os critérios de acordo com as métricas recuperadas a partir do *Repositório de Critérios de Qualidade*, como mostrado na Tabela 2.

Tabela 2. Repositório de Critérios de Qualidade

[1] http://www.aktors.org/publications/ontology/portal

[2] http://ieee.rkbexplorer.com/sparql/

[3] http://acm.rkbexplorer.com/sparql/

[4] http://dblp.rkbexplorer.com/sparql/

Critérios	Métricas	Comentários
Disponibilidade	SELECT * WHERE {?S ?P ?O} limit 1	Verifica se a fonte responde a uma consulta SPARQL. Caso positivo, a fonte está disponível (*um*). Caso ultrapasse um tempo limite de espera, a fonte não está disponível (*zero*).
Completude do Esquema	$CE = \sum C_f / \sum C_{app}$	Soma dos conceitos existentes na fonte dividido pela soma dos conceitos do conjunto de consultas da aplicação.

A Tabela 3, por sua vez, apresenta os valores dos critérios para as três fontes consideradas. No exemplo, considera-se que a fonte de dados DBLP não se encontra disponível no ato da consulta. Para o cálculo da completude, é necessário identificar os conceitos presentes nas consultas da aplicação. Conceitos podem ser identificados como os principais termos que descrevem uma consulta, os quais podem ser extraídos dos predicados presentes nos padrões de triplas da consulta. Neste caso, tem-se que: C_{Q1} = {*title, author, full-name*}, C_{Q2} = {*title, author, full-name, date, journal*} e C_{Q3} = {*title, author, full-name, date*}. Além disso, criamos um C_{app} com todos os conceitos do conjunto de consultas da aplicação tal que, consultando a descrição das fontes de dados encontram-se as seguintes intercessões de conceitos entre o conjunto de consultas da aplicação e as fontes: C_{ACM} = {*title, author, full-name, date*}, C_{IEEE} = {*title, author, date*}, C_{DBLP} = {*title, author, full-name, date, journal*}. Assim, na Tabela 3, tem-se que a fonte de dados ACM, por exemplo, tem valor de completude igual a 0,8 (*CE*= 4/5), uma vez que possui *quatro* dos *cinco* critérios relevantes para a aplicação.

Tabela 3. Valores de Critérios e Classificação das Fontes

Critério \ Fonte	ACM	IEEE	DBLP
Disponibilidade	1	1	0
Completude do Esquema	0,8	0,6	1
Classificação	**1** (0,82)	**2** (0,74)	**3** (0,40)

Por fim, a *Medida de Qualidade das Fontes de Dados* é calculada para cada fonte, usando a fórmula seguinte:

*MQ = ((critério 1 * peso 1) + (critério 2 * peso 2) + ... + (critério n * peso n)) / n*

Para a fonte de dados IEEE, por exemplo, tem-se a seguinte medida de qualidade: *MQ* = (1 * 100%) + (0,6 * 80%) / 2 = 0,74. O resultado da avaliação **(Medida de Qualidade das Fontes de Dados)** é mostrado em ordem crescente segundo a qualidade da fonte: **MQ = {(0,82 – ACM), (0,74 – IEEE), (0,40 – DBLP)}.** Como resultado conclui-se que o conjunto de dados ACM possui uma qualidade melhor de acordo com os requisitos da aplicação. Apesar do DBLP possuir uma completude do esquema de 100%, o fato de não estar disponível no momento da avaliação contribuiu negativamente para sua classificação final.

4. Trabalhos Relacionados

Em 2012, Mendes *et al*. (2012) simplificaram a tarefa de consumir dados de fontes *linked data* de alta qualidade por meio do *framework* SIEVE. A tarefa de avaliação de qualidade do SIEVE é realizada por um módulo flexível onde o usuário pode escolher quais características dos dados podem ser indicativos de uma boa qualidade, como essa qualidade pode ser quantificada e como ela deve ser armazenada no sistema. O SIEVE utiliza pelo menos *três* critérios iniciais para melhorar a qualidade dos dados: completude, concisão e consistência. O uso desses critérios permite o SIEVE remover valores conflitantes e redundantes.

Cordeiro *et al.* (2011) propuseram uma arquitetura para gerenciar e enriquecer a semântica ao publicar dados governamentais em formato *Linked Data*. O principal objetivo da proposta é propor uma plataforma que oferece suporte à exposição,

compartilhamento e associação dos recursos em formato *Linked Data* por meio de um ambiente amigável ao usuário. No processo de conversão dos dados para o formato *Linked Data* todas as etapas são monitoradas de forma a garantir a proveniência (critério de qualidade) e enriquecer a semântica da informação.

A proposta apresentada neste trabalho se diferencia das anteriores por ser uma arquitetura passível de ser expandida em fontes de dados na Web e não apenas em fontes que seguem os princípios *Linked Data*. Além disso, o repositório extensível de critérios de qualidade permite uma maior flexibilidade com relação aos domínios de aplicação e fontes de dados que podem ser considerados.

5. Avaliação dos Resultados e Estado Atual do Trabalho

Como principais atividades realizadas até o momento destacam-se o levantamento do referencial bibliográfico sobre critérios de qualidade para avaliação de qualidade de fontes *linked data* e a especificação da arquitetura proposta para avaliação da qualidade de fontes *linked data* com seus principais componentes. Para validar a proposta, um protótipo está sendo implementado com todos os componentes descritos na Figura 1. Atualmente, o foco do trabalho está na implementação do *Editor de Requisitos de Qualidade* que irá disponibilizar os critérios de qualidade para um especialista no domínio da aplicação. Como trabalho futuro tem-se a possibilidade de generalização da arquitetura para avaliação de fontes de dados na Web.

Referências

Batista, M. C. M. (2008)"Schema Quality Analysis in a Data Integration System". PhD Thesis, Universidade Federal de Pernambuco, 2008.

Bizer C., Heath T., Berners-Lee T. (2009) "Linked data – the story so far", In: Int. J. Semantic Web Inf. Syst.

Cordeiro, K.F., Faria, F.F., Pereira, B.O., Freitas, A., Ribeiro, C.E., Freitas, J.V.V.B., Bringuente, A.C., Arantes, L.O., Calhau, R., Zamborlini, V., Campos, M.L.M., and Guizzardi, G. (2011) "An approach for managing and semantically enriching the publication of Linked Open Governmental Data", In: Proceedings of the 3rd Workshop in Applied Computing for Electronic Government (WCGE), pages 82-95.

Flemming, A. (2010) "Quality characteristics of linked data publishing data sources", Master's thesis, Humboldt-Universitätzu Berlin.

Hogan, A., Harth, A., Passant, A., Decker, S., and Polleres, A. (2010) "Weaving the pedantic web", In: LDOW 2010.

Lóscio, B. F.; Batista, M. C. M.; Souza, D.; Salgado, A. C. (2012) "Using Information Quality for the Identification of Relevant Web Data Sources: A Proposal", In: iiWAS 2012.

Mendes, P., Mühleisen, H., and Bizer, C. (2012) "Sieve: Linked data quality assessment and fusion", In: Proceedings of LWDM (March 2012).

W3C (2013). Disponível em <http://www.w3.org/RDF/>. Acesso em: Agosto de 2013.

Wang, R. Y. and Strong, D. M. (1996) "Beyond accuracy: What data quality means to data consumers", In: Journal on Management of Information Systems, 12(4):5-34.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auear, A. (2012) "Quality Assessment Methodologies for Linked Open Data", In: IOS Press 2012.