Registro de procedência de ligações RDF em Dados Ligados

Jonas F. S. M. De La Cerda[1], **Maria Cláudia Cavalcanti**[1]

[1]Instituto Militar de Engenharia
Praça General Tibúrcio, 80 – Praia Vermelha – Rio de Janeiro – RJ

Abstract. *As many tools have been created to support linked data consumption and publishing, there is a demand for quality assessment and to verify these data. To make this possible, data about this consumption should be recorded. This paper presents an extension to a framework with the goal to support the recording and publishing of the information about the creation and consumption of linked data, in order to provide input for later quality assessment.*

Resumo. *Com a criação de ferramentas para consumir, relacionar e publicar dados ligados, surge a demanda para avaliar e comprovar a qualidade destes dados. Para tal, é necessário que informações sobre este consumo sejam registradas. Este trabalho propõe a extensão de uma arquitetura a fim de suportar o registro e publicação de informações sobre a criação destes dados, a fim de prover insumos para posterior avaliação.*

1. Introdução

Com o desenvolvimento e adoção da *web* semântica, vieram padrões e formatos para integrar dados e informações oriundos de diferentes fontes. Há iniciativas para disponibilizar dados em formatos padronizados, para que estes possam ser consumidos (e relacionados) com dados de diferentes fontes. Uma destas iniciativas é o *Linked Data* (dados ligados) [1], que consiste em interligar dados de diversas fontes segundo alguns princípios. Estes princípios são: disponibilizar os dados em um formato padronizado – no caso o RDF (*Resource Description Framework*) [2] – e fornecer meios para acessar e identificar os dados disponibilizados.

É possível criar aplicações mais ricas em informação através do consumo dos dados e seus relacionamentos de diversas fontes. Para tal, é necessário considerar problemas como a obtenção do dado, mapeamento de esquemas e vocabulários, e análise de qualidade do dado. Diante destes problemas, diversas ferramentas foram criadas para facilitar a integração e consumo dos dados ligados, algumas listadas em [Bizer *et al.* 2009]. Não há a preocupação em registrar informações de como estas novas relações foram geradas, criando um problema para provar a confiabilidade e corretude do processo empregado.

Este trabalho propõe uma arquitetura a fim de suportar o registro de informações sobre a criação das interligações de recursos RDF, ou seja, registrar as informações de quais processos foram utilizados para criação, quais parâmetros configuraram estes processos, quais os resultados destes processos. Acredita-se que tais informações podem ajudar em futura análise de qualidade dos dados, tornando-se um ativo tanto para quem con-

[1]http://www.w3.org/DesignIssues/LinkedData.html
[2]http://www.w3.org/TR/REC-rdf-syntax/

some os dados quanto para quem os que publica. A seção 2 deste artigo apresenta os conceitos básicos de dados ligados. A seção 3 apresenta trabalhos relacionados, constando de: uma arquitetura prévia e sua implementação, e modelos de dados de procedência. A seção 4 apresenta a arquitetura proposta, e a seção 5 apresenta as conclusões e extensões do projeto.

2. Dados Ligados

Uma vez que consumir e integrar estes dados se dá de forma mais flexível, é possível escapar do contexto de uma *web* ultrapassada onde aplicações devem prever o consumo de fontes de dados previamente definidos, criando uma *web* onde a informação provida por aplicações pode evoluir ao longo do tempo, junto com o surgimento de novas fontes de dados. Para tirar proveito dos dados ligados, Berners-Lee elucida em um documento [3] regras para publicar (e consumir) os dados ligados: usar URIs válidas para nomear seus recursos (dados, coisas, entidades, etc), de forma que agentes (pessoas ou sistemas) recebam informações úteis – preferencialmente em formato inteligível – ao acessar tais endereços, e, principalmente incluir ligações (links) para recursos em outras fontes de dados, para que novos conhecimentos possam ser descobertos.

Em um tutorial [4] feito por Bizer, define-se uma ligação RDF como uma tripla no formato "sujeito - predicado - objeto" onde o sujeito é ligado ao objeto através de um predicado. As ligações RDF onde o sujeito está em um conjunto de dados e o objeto está em um conjunto de dados distinto são chamados de ligações externas.

3. Trabalhos Relacionados

Existem diversas aplicações utilizando dados ligados. Tais aplicações vão desde *endpoints* SPARQL – formulários onde insere-se uma consulta em SPARQL e recebe-se o resultado da consulta, usualmente no formato de alguma serialização RDF – até aplicações mais complexas como os *websites* da BBC. Em [Kobilarov *et al.* 2009] são apresentados os mecanismos utilizados por estes sistemas a fim de consumir e gerar ligações com outros provedores de dados ligados. São explorados os mecanismos utilizados para interligar os diversos sistemas (legados e atuais) da BBC à nuvem do movimento *Linking Open Data* [5], os mecanismos para reutilização e redirecionamento para conteúdos de outros provedores de dados, os mecanismos da publicação de dados dos programas da emissora.

Em [Bizer *et al.* 2009] é identificada uma arquitetura comum de aplicações voltadas para dados ligados. Tal arquitetura é ilustrada na Figura 1, adaptada de [Isele *et al.* 2010], excluindo-se a parte tracejada da figura, que representa um coletor de dados de procedência a ser explicado mais adiante. Para consumir – importar, associar e publicar – os dados ligados da *web*, uma aplicação tem que considerar problemas como obtenção do dado, mapeamento de esquemas e vocabulários e análise de qualidade do dado. Existe uma implementação funcional de um arcabouço para executar todas as etapas da integração dos dados ligados previstas pela arquitetura comum, o LDIF (*Linked Data Integration Framework*) [Schultz *et al.* 2011].

[3] http://www.w3.org/DesignIssues/LinkedData.html

[4] http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/

[5] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

Ao passo que o LDIF ataca os problemas de mapeamento de esquemas e vocabulários, resolução de identidades, importação, publicação e descoberta de ligações (relações entre recursos), o arcabouço se apresenta deficiente no quesito da procedência dos dados. Procedência refere-se à linhagem dos dados, isto é, as origens e histórico de processamento de objetos e processos [Bose e Frew 2005], ou seja, a procedência possui um papel importante em evidenciar a qualidade dos dados gerados.

A deficiência do LDIF quanto à captura da procedência é evidente pois os únicos dados de procedência publicados são os dados relativos à importação inicial dos dados, ou seja, qual a origem dos dados importados. Dados importantes de procedência como a parametrização de processos de similaridade sintática e semântica, resultados da execução de processos, dentre outros, não são contemplados, nem pelo LDIF e nem pela arquitetura de aplicações de dados ligados.

Os dados de procedência podem servir de insumo para análise de qualidade dos dados gerados. Pode-se atribuir maior confiabilidade a dados gerados por processos que foram configurados com limites mais restritos. Por exemplo, é possível atribuir maior confiabilidade às ligações geradas por processos de cálculo de similaridade que tenham sido configurados com um limite de similaridade maior que 0.95 (95%). Em [Mendes *et al.* 2012] são ilustrados tanto exemplos de avaliação de qualidade dos dados quanto de fusão de dados. Um dos exemplos mostrados por Mendes, é a atribuição de reputação aos dados de acordo com sua origem, e, a pontuação (*scoring*) de acordo com o quão recente o dado é.

Dada a importância dos dados de procedência, alguns modelos influenciaram este trabalho. O mais notável é o OPM (*Open Provenance Model*) [Moreau *et al.* 2011], que descreve as relações causais e de dependência entre artefato (que representa o estado imutável de um objeto), processo (que representa ações efetuadas em um artefato, ou causadas por) e agente (que representa entidades que podem facilitar, controlar ou influenciar um processo de alguma forma). Os outros modelos que influenciaram este são o *Provenir* [Sahoo e Sheth 2009] e o PROV-DM [6]. Os conceitos definidos pelo OPM estão presentes também nestes modelos. No caso do *Provenir*, estes conceitos são mais especializados (e.g. diferenciação de dados e parâmetros). Já o PROV-DM não é tão específico quanto aos artefatos, porém possui muitas definições das relações de dependência e causalidade, inclusive sendo especificadas formalmente.

4. Arquitetura Proposta

Este trabalho propõe que a arquitetura das aplicações ainda deficiente na questão da procedência de dados contemple tal aspecto, fornecendo um modelo de dados para o processo de integração de dados ligados. A arquitetura deve contemplar o aspecto de procedência em todas as etapas dos processos de consumo e integração, conforme mostra a Figura 1. Para tal, diversos modelos de procedência devem ser estudados, a fim de definir um modelo que seja compatível com os modelos já existentes e difundidos.

O modelo de procedência a ser adotado na nova arquitetura deve não somente contemplar a diferenciação entre dados e parâmetros, mas também deve diferenciar os processos empregados na integração dos dados ligados, considerando a hierarquia

[6] http://www.w3.org/TR/prov-dm/

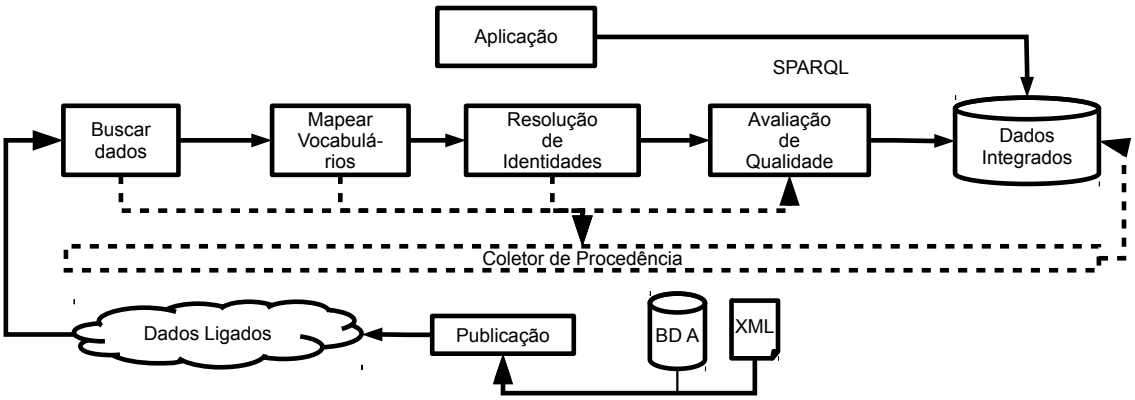

Figura 1. Arquitetura de aplicações consumidoras de dados ligados considerando os aspectos de procedência de dados.

de técnicas empregadas tanto no mapeamento de vocabulários quanto na descoberta de links. Uma visão de como essas técnicas podem ser classificadas foi apresentada por [Euzenat e Shvaiko 2007] e foram também estudadas por [Silva 2010], que relacionou esta visão com as medidas de similaridades definidas por [Ehrig 2007].

Até o momento, o modelo considera alguns aspectos básicos quanto aos tipos de processos utilizados na integração e consumo de dados ligados, e, considera uma categorização dos dados em questão. Os tipos de processo contemplados até o momento são processos de importação – processos que obtém os dados de seus provedores originais – e processos de integração. Os processos de integração se encontram categorizados como processos de mapeamento (de vocabulários) e processos de ligação.

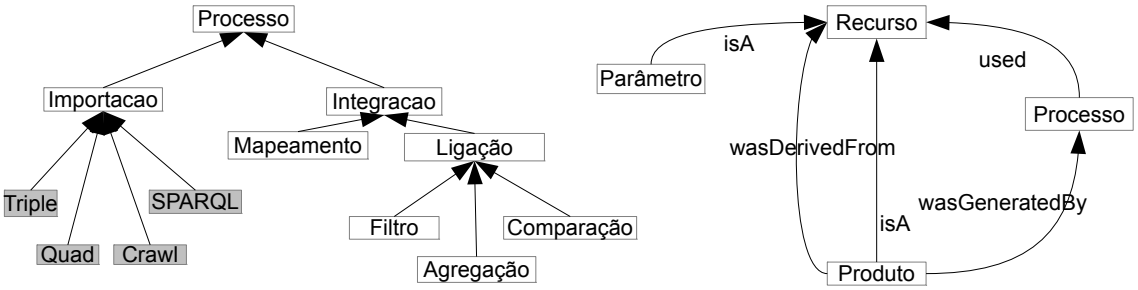

Figura 2. Conceitos do modelo de dados de procedência.

Os processos de mapeamento tratam-se de definições de pareamentos de conceitos de um vocabulário a outro, como parear foaf:Person e dbpedia:Person ou foaf:name e rdfs:label. Os processos de ligação tratam-se de execuções de processos que geram ligações RDF através de alguma computação. Tais processos podem ser processos de agregação – como médias, máximos, mínimos – processos de filtragem – como selecionar recursos que possuam uma determinada propriedade dentro de um intervalo de valores – e processos de comparação – como comparar rótulos RDF através de distância de edição, ou comparar a categorização de dois recursos.

Pode-se dizer que há uma equivalência entre os conceitos de processo do OPM e processo do modelo proposto. Uma ideia inicial do modelo é ilustrada pela Figura 2,

onde os conceitos em cinza-escuro representam extensões dos processos de importação, inclusive já implementados no LDIF.

No que concerne ao conceito de artefato do OPM, há uma relação de equivalência com o conceito de recurso, subcategorizado em parâmetro e produto, como mostra a Figura 2. A diferença entre produtos e parâmetros é que produtos são gerados por processos, ou seja, para gerar cada produto foram consumidos tempo e recursos computacionais.

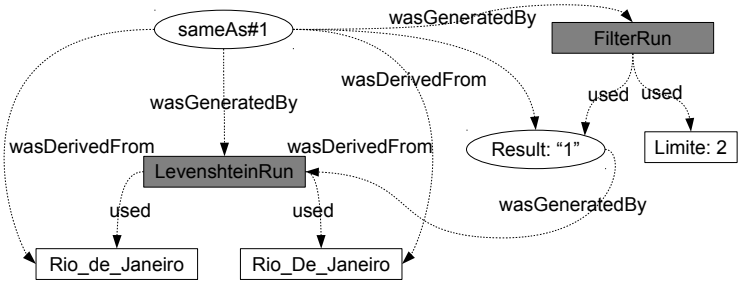

Figura 3. Exemplo de aplicação de modelo.

A Figura 3 exemplifica uma aplicação bastante básica do modelo, a criação de uma ligação do tipo "owl:sameAs" entre dois recursos de rótulos "Rio_de_Janeiro" e "Rio_De_Janeiro", respectivamente. A geração da ligação se dá em dois passos, o primeiro sendo a comparação entre os rótulos dos recursos através de um algoritmo que calcula distância de edição entre duas cadeias de caracteres e o segundo filtrando apenas os produtos que tenham sido gerados com distância de edição abaixo de 2. Na Figura 3, os produtos estão representados por elipses, os parâmetros por retângulos claros e os processos por retângulos escuros. Explicitar todas as relações causais entre dados e processos pode gerar um excesso de informações, que é problema conhecido e já foi discutido em [Heinis e Alonso 2008], não sendo o foco deste trabalho.

Em resumo, modelo e arquitetura propostos encapsulam os executores dos processos envolvidos em cada etapa do fluxo da integração e consumo de dados ligados, a fim de registrar e representar os dados de procedência de acordo com a natureza dos processos envolvidos na criação das ligações RDF entre recursos, bem como a natureza dos parâmetros que configuram estes processos e resultados destes processos. Dessa forma, esses dados de procedência passam a estar disponíveis para um usuário avaliar confiabilidade e autenticidade das ligações geradas, avaliar a qualidade e efetuar fusão de dados ligados – como é o caso do Sieve [Mendes *et al.* 2012] – e reproduzir o processo de geração de ligações RDF.

5. Conclusão

Este artigo apresenta uma proposta para o problema do registro e representação de procedência de dados na atividade de integração e consumo de dados ligados. A sua principal contribuição é a extensão de modelos de procedência já estabelecidos e ainda em definição, adaptando-os para registrar informações mais específicas sobre o consumo e integração de dados ligados. A partir de uma arquitetura já existente – o LDIF – de código aberto, estende-se sua funcionalidade de modo a suportar o registro dessas informações. No momento a extensão proposta está em fase de implementação. O modelo de dados

proposto ainda passa por refinamentos, devendo evoluir a fim de especificar os processos envolvidos e tipos de dados e parâmetros.

Trabalhos futuros incluem o estabelecimento de políticas de descarte e seleção de ligações RDF, com base nos dados de procedência disponibilizados. Além disso, conforme as ligações RDF são rastreadas e associadas às informações de procedência, é possível estabelecer e configurar mecanismos de inferência baseados nessas informações.

Acknowledgements

The authors would like to thank CNPq (309307/2009-0; 486157/2011-3) and FAPERJ (E-26/111.147/2011) for partially funding their research projects.

Referências

Bizer, C., Heath, T., e Berners-Lee, T. (2009). Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22.

Bose, R. e Frew, J. (2005). Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys*, 37:1–28.

Ehrig, M. (2007). *Ontology Alignment: Bridging the Semantic Gap*, volume 4 of *Semantic Web And Beyond Computing for Human Experience*. Springer.

Euzenat, J. e Shvaiko, P. (2007). *Ontology matching*. Springer.

Heinis, T. e Alonso, G. (2008). Efficient lineage tracking for scientific workflows. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, SIGMOD '08, pages 1007–1018, New York, NY, USA. ACM.

Isele, R., Jentzsch, A., e Bizer, C. (2010). Silk server - adding missing links while consuming linked data. In *1st International Workshop on Consuming Linked Data (COLD 2010), Shanghai*.

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., e Lee, R. (2009). Media meets semantic web — how the bbc uses dbpedia and linked data to make connections. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009 Heraklion, pages 723–737, Berlin, Heidelberg. Springer-Verlag.

Mendes, P. N., Mühleisen, H., e Bizer, C. (2012). Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, EDBT-ICDT '12, pages 116–123, New York, NY, USA. ACM.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P. T., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E. G., e den Bussche, J. V. (2011). The open provenance model core specification (v1.1). *Future Generation Comp. Syst.*, 27(6):743–756.

Sahoo, S. S. e Sheth, A. (2009). Provenir ontology: Towards a framework for escience provenance management. Microsoft eScience Workshop.

Schultz, A., Matteini, A., Isele, R., Bizer, C., e Becker, C. (2011). *LDIF - Linked Data Integration Framework*, pages 1–4.

Silva, V. d. S. (2010). Uma abordagem para alinhamento de ontologias biomédicas para apoiar a anotação genômica. Master's thesis, Universidade Federal do Rio de Janeiro.