Linked Geospatial Data: desafios e oportunidades de pesquisa

Tiago Henrique V. M. de Moura, Clodoveu A. Davis Jr.

Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG - Brasil

{thvmm,clodoveu}@dcc.ufmg.br

Abstract. *Linked Geospatial Data is a proposal for the dissemination of geographic information in an open format, based on standards adopted by the Web of Data. The semantic enrichment that comes from knowledge about relationships among spatial data is a potential source of solutions for traditional problems in geographic information retrieval, such as ambiguity resolution and the recognition of the geographic context of documents. This work discusses the applicability of linked data concepts to these problems and presents a set of challenges and opportunities for research on geographic information retrieval and related topics.*

Resumo. *Linked Geospatial Data é uma proposta para a disseminação de informações geográficas em um formato aberto, baseado em padrões da Web of Data. O enriquecimento semântico decorrente do conhecimento sobre relacionamentos entre dados geoespaciais é uma fonte potencial de soluções para problemas tradicionais em recuperação de informação geográfica, tais como resolução de ambiguidade e reconhecimento do contexto espacial de documentos. Este trabalho discute a aplicabilidade dos conceitos de* linked data *a esses problemas e apresenta uma lista de desafios e oportunidades para a pesquisa sobre recuperação de informação geográfica e tópicos correlatos.*

1. Introdução

Documentos contendo texto não estruturado são bastante comuns na Web. Esses documentos são interligados por *hyperlinks*, que podem ser interpretados como relacionamentos entre eles (Bizer *et al*. 2009). Dados podem ser organizados usando esse mesmo mecanismo de relacionamento adotado entre documentos na Web. Isso pode ajudar a enriquecê-los, tornando mais simples o processamento por máquinas, e também facilitando a integração entre diversas fontes. Essa ideia é a base da proposta denominada *Linked Data* (Berners-Lee 2006).

Dentre os dados que podem ser organizados como *linked data*, destacam-se os dados geoespaciais, por sua importância e potencial de integração. Seu uso enriquece as aplicações que envolvem dados georreferenciáveis, mas existem problemas, como a falta de fontes confiáveis e a ambiguidade de nomes de lugares. Alguns desses problemas são abordados na literatura, e soluções propostas envolvem recursos como dicionários toponímicos (*gazetteers*) e dados de contribuições voluntárias, criados e mantidos em projetos isolados. A integração dessas fontes de dados por meio dos conceitos de *linked data* no contexto geoespacial tem grande impacto potencial sobre as aplicações.

Este trabalho apresenta conceitos sobre *linked data* geoespaciais, e aborda desafios e questões para investigação sobre sua utilização em recuperação de informação geográfica. Estamos especificamente interessados no potencial suporte de *linked data*

geoespacial a problemas referentes ao reconhecimento do contexto geográfico em documentos da Web, em que frequentemente são utilizados dados de referência, como *gazetteers*. O restante do artigo está organizado da seguinte forma. A Seção 2 aborda os conceitos básicos de *linked data*. Na Seção 3 esses conceitos são expandidos para o contexto geoespacial. A Seção 4 descreve os desafios que surgem nesta nova ótica. Finalmente, as conclusões deste estudo e trabalhos futuros são apresentados na Seção 5.

2. *Linked Data*

O paradigma *linked data* pode ser definido de maneira bastante ampla como a utilização de protocolos da Web para criação de relacionamentos tipados entre diferentes fontes de dados. O objetivo é criar uma fonte de dados integrada e aberta, denominada *Web of Data*. Para que esta ideia seja aplicável, algumas premissas para publicação de dados foram estabelecidas (Berners-Lee 2006): (1) os dados precisam obedecer ao padrão RDF (*Resource Description Framework*); (2) os objetos precisam ser identificados por URIs (*Universal Resource Identifier*); (3) os dados devem estar acessíveis via protocolo HTTP; e (4) os objetos devem ser referenciados através de suas URIs.

Os dados precisam estar formatados em um padrão legível por máquinas, daí a escolha do RDF. Nesse formato, os dados são descritos formando triplas `<objetoA, predicado, objetoB>`. Assim, os objetos A e B, identificados por suas URIs, se relacionam de maneira explícita pelo predicado que compõe a tripla. Isso implica no uso de HTTP (premissa 3) para efetivar o relacionamento. A utilização de URIs como identificadores de objetos tenta simplificar a maneira de acessá-los, já que basta navegar até a URI para acessar os dados de determinado objeto. A premissa 4 é importante para interconectar as fontes de dados, criando contexto entre elas e facilitando a navegação entre os elementos que compõem a *Web of Data*.

Um exemplo de tripla RDF é indicado na Figura 1. A primeira entidade (primeira linha) representa Paris, e a segunda a França (terceira linha). O relacionamento (segunda linha) indica que Paris é uma cidade francesa.

```
<http://sws.geonames.org/2988507/,
        gn:parentCountry,
 http://sws.geonames.org/3017382/>
```

Figura 1. Exemplo de tripla RDF

Um grupo formado em janeiro de 2007 é o responsável por manter o principal projeto de criação de *linked data*, o *Linking Open Data* (LoD). O ponto de partida foi a base DBpedia, que conta com informações sobre alguns tópicos da Wikipedia (Bizer *et al*. 2009) e inicialmente integrava 12 conjuntos de dados. O projeto, que inicialmente era conduzido basicamente por pesquisadores e desenvolvedores de pequenas empresas, hoje atrai mais atenção e contava com quase 300 conjuntos de dados em 2011.

Em setembro de 2011 a *Web of Data* tinha mais de 31,5 bilhões de triplas (Tabela 1). No entanto, o número de relacionamentos entre bases de dados (*outlinks*) é relativamente baixo, pouco mais que 500 milhões de links. Isso indica que faltam referências a objetos externos, refletindo uma dificuldade na aplicação da premissa 4 da *Web of Data*. Também indica que os dados de cada fonte não estão sendo suficientemente enriquecidos pela integração a outras fontes, ou que os conjuntos de dados conteriam descrições

muito detalhadas de seus objetos, fazendo o número de triplas crescer sem o correspondente aumento nos *outlinks*.

Tabela 1. Números da Web of Data em setembro de 2011[1]

Domínio	Número de *datasets*	Triplas	%	Outlinks	%
Mídia	25	1.841.852.061	5,82 %	50.440.705	10,01 %
Geográfico	31	6.145.532.484	19,43 %	35.812.328	7,11 %
Governamental	49	13.315.009.400	42,09 %	19.343.519	3,84 %
Publicações	87	2.950.720.693	9,33 %	139.925.218	27,76 %
Múltiplos domínios	41	4.184.635.715	13,23 %	63.183.065	12,54 %
Ciências da vida	41	3.036.336.004	9,60 %	191.844.090	38,06 %
Conteúdo gerado por usuários	20	134.127.413	0,42 %	3.449.143	0,68 %
TOTAL	295	31.634.213.770		503.998.829	

3. *Linked Data* Geoespaciais: integração de fontes de dados de referência

Dados geoespaciais são comumente utilizados em sistemas de informação geográficos (SIG), mas nem sempre é fácil ou possível extraí-los para uso em outras aplicações, ou mesmo em outros SIG. *Linked data* possibilita a reutilização desses dados, facilitando o surgimento de novas aplicações e inovações.

Além das novas possibilidades criadas por *linked data*, existem desafios de recuperação de informação geográfica nos quais esse conceito pode ser muito útil. Uma classe importante de problemas envolve o reconhecimento do contexto geográfico de documentos. Nesses problemas, é necessário reconhecer nomes de lugares e outras referências espaciais no texto, mas frequentemente ocorre ambiguidade, ou seja, termos usados como nome de um lugar podem se referir a outras entidades ou a lugares homônimos. *Gazetteers* são usados como fontes de nomes válidos de lugares, permitindo o reconhecimento. Para a desambiguação, no entanto, em geral mais informação é necessária. Em um trabalho anterior (Machado et al. 2010), um gazetteer enriquecido com informação semântica foi proposto, e foi demonstrada uma técnica de desambiguação que explora o relacionamento espacial e semântico entre lugares (Machado et al. 2011).

A utilização de *linked data* nesse contexto traz diversas vantagens, relacionadas à expansão do conhecimento que se tem sobre um objeto e seus relacionamentos. Como na *Web of Data* todos os relacionamentos entre objetos são semanticamente bem definidos, estes poderiam ser utilizados no algoritmo proposto por Machado *et al*. 2011.

O *gazetteer* GeoNames, uma das primeiras fontes de dados do projeto LoD, contém atualmente mais de 8,3 milhões de referências. Tem cobertura global, porém seu nível de detalhamento é heterogêneo, não inclui muitos dados intraurbanos, e os relacionamentos entre lugares estão restritos a hierarquias de subdivisão espacial e a algumas relações de vizinhança. O *gazetteer* proposto por Machado et al. (2010), por outro lado, traz detalhamento urbano em algumas cidades brasileiras e relacionamentos semanticamente mais ricos, porém seu escopo é limitado ao Brasil. A interligação dessas fontes de

[1] http://lod-cloud.net/state/

dados, usando as técnicas de *linked data*, resultaria em uma base de conhecimento mais ampla, com melhores possibilidades de uso em problemas de desambiguação. Porém, essa interligação só pode ser realizada com segurança caso seja possível garantir a correspondência entre entidades das bases envolvidas – o que é também um problema de desambiguação e casamento de registros (*records matching*). Essa dificuldade faz parte dos desafios enfrentados atualmente para a expansão das ligações entre bases de *linked data*, no caminho de realizar todo o potencial dessa ideia. A próxima seção discute os principais desafios que identificamos para *linked data* geoespaciais

4. Desafios

A disponibilidade de grandes volumes de *linked data* geoespaciais pode ser importante para a pesquisa em recuperação de informação geográfica e para o reconhecimento de lugares em texto, a partir da integração de fontes de referência. Os tópicos a seguir descrevem alguns desafios de pesquisa nessa direção.

Manutenção e atualização de dados de *gazetteers*: contribuição voluntária. A manutenção e atualização dos dados contidos em *gazetteers* são tarefas muito custosas, principalmente se alguma intervenção manual é necessária. Uma solução para este problema é a utilização de contribuições voluntárias (conhecidas como *Volunteered Geographic Information*, VGI) (Goodchild, 2007a; McDougall, 2009). Isso é feito em algumas aplicações na Web, como o Wikimapia[2], mas é necessário integrar dados coletados por essas aplicações à *Web of Data*. Dados informados voluntariamente podem também ser usados para aplicações que funcionam em tempo real, como o monitoramento do trânsito urbano (Davis Jr *et al*. 2011). VGI associada às diretrizes do *linked data* fornece um forte alicerce para construção de sistemas capazes de resolver consultas complexas com bastante qualidade (Keßler *et al*. 2009), já que associaria o conhecimento dos usuários aos dados já existentes de uma maneira semanticamente mais rica. Porém, observe-se que dados gerados por usuários representam ainda menos de 1% das triplas existentes na *Web of Data*.

Resolução de entidades e desambiguação. Em todo tipo de aplicação ou problema de pesquisa em que são usados nomes de lugares, existem problemas ligados à desambiguação. Existe a ambiguidade *geo*/*geo*, ou seja, quando mais de um lugar tem o mesmo nome, e *geo*/*não-geo*, quando lugares usam nomes também usados para outras entidades (Amitay *et al*. 2004). Pode-se usar *gazetteers* no reconhecimento de possíveis referências a lugares, mas para resolver a ambiguidade é necessário analisar outros fatores, como a coocorrência de nomes de lugares relacionados ou a presença de termos fortemente associados a lugares citados no texto. Os *gazetteers* tradicionais, como o GeoNames, não são uma boa fonte para esses relacionamentos, porém a integração deles a fontes como a DBPedia e a *gazetteers* com conteúdo intraurbano pode prover elementos para resolver o problema. Este é um ponto em que *linked data* podem oferecer boas soluções para ampliar o conteúdo das bases de referência, dando clareza para caracterizar os lugares citados, e para prover recursos que permitam caracterizar semanticamente relacionamentos expressos nas triplas RDF. Por outro lado, a resolução de entidades e a desambiguação de nomes são importantes também para promover maior integração entre fontes na *Web of Data*.

[2] http://wikimapia.org/

***Data Fusion* e redundância.** A redundância na *Web of Data* é um problema a partir do momento em que informações sobre a mesma entidade do mundo real ocorrem em diferentes conjuntos de dados, sendo que em cada um foi criada uma URI diferente. Existe um tipo de relacionamento na *Web of Data* que caracteriza dois objetos como iguais. Entretanto, é difícil estabelecer quando os objetos de fato coincidem, pois mesmo que correspondam à mesma entidade do mundo real, podem não compartilhar a mesma representação (Jain *et al*. 2010), por exemplo variando a escala. Nesse caso, como integrar as representações? Esse exemplo ilustra a dificuldade em se conseguir aumentar, de forma segura, a quantidade de conexões entre fontes de dados. A criação de mecanismos que reduzam a redundância na *Web of Data* e melhorem a qualidade das representações é importante para o futuro das pesquisas relacionadas a *linked data*.

Dados Temporais. A utilização de triplas RDF é capaz de ajudar a estabelecer a semântica dos relacionamentos, mas existem limitações importantes, principalmente na representação de dados temporais (Erwig *et al*. 1999). No GeoNames, por exemplo, cidades possuem um atributo referente a população, mas não existe uma data de referência para esse valor, logo o dado se torna incompleto e, dependendo do estudo que está sendo feito, até mesmo inútil. Pode ser necessário acrescentar um quarto atributo às triplas para estabelecer uma referência temporal para os dados. Um exemplo representativo da necessidade de se acrescentar atributos temporais a conjuntos de dados geoespaciais é a associação de dados censitários aos polígonos dos municípios brasileiros. Como as fronteiras municipais mudam ao longo do tempo, por exemplo no evento do desmembramento de um município, existe dificuldade em modelar e implementar uma série histórica geoespacial de dados demográficos pensando em *linked data*.

Qualidade, relevância e confiança. Os dados na *Web of Data* podem ser vistos sob duas perspectivas: (1) os *datasets* isolados e (2) os objetos que compõem cada um dos *datasets*. Em ambas, a existência de métricas para avaliação de qualidade, relevância e confiança é muito importante. Cada conjunto de dados tende a abordar um nicho específico, e o próprio LoD categoriza seus *datasets*. Dessa forma, é importante criar mecanismos confiáveis de categorização para que uma nova aplicação possa escolher os conjuntos de dados que mais se aproximem de suas necessidades. Aplicações que operam sobre toda a *Web of Data*, como por exemplo mecanismos de busca, precisariam estimar a relevância dos objetos relacionados a uma consulta, estabelecendo um ranqueamento. Uma métrica que poderia ser adaptada é o PageRank (Brin and Page 1998), mas no lugar de documentos teria que ser estimada a relevância das bases, ou até mesmo dos objetos que a compõem, diante de um perfil de usuário ou aplicação.

5. Conclusões e trabalhos futuros

Linked data é um novo paradigma para a integração de dados, simplificando os esquemas de relacionamento e enriquecendo a semântica. Embora sua utilização pareça vantajosa, a tecnologia é nova e existem diversos desafios a serem explorados e áreas de atuação a serem descobertas. Este artigo representa um trabalho de pesquisa em andamento, cujo objetivo final é verificar a aplicabilidade de *linked data* geoespaciais a problemas de recuperação de informação geográfica como alternativa à construção de bases de referência. Nos interessa avaliar o compromisso entre os ganhos decorrentes da integração de fontes diversas de dados de referência e os ganhos para o processo de recuperação de informação. Nesse sentido, os desafios apresentados constituem ao mesmo tempo barreiras para a construção imediata de aplicações e oportunidades para explora-

ção de novos conceitos. Pretende-se, inicialmente, promover a integração entre o Onto-Gazetteer (Machado *et al*. 2011) e o GeoNames, e avaliar o impacto dessa integração para as aplicações em recuperação de informação geográfica que usam bases de referência. Outras atividades incluem a avaliação das linguagens de consulta propostas para *linked data* e a continuação de esforços anteriores de pesquisa no sentido da detecção de tópicos na Wikipedia em integração com os *gazetteers* (Alencar and Davis Jr 2011), de modo a obter listas de termos *não-geo* associados a lugares.

Agradecimentos

O presente trabalho foi parcialmente financiado com recursos dos projetos CEX-PPM-00518/13 (FAPEMIG), 560027/2010-9 e 308678/2012-5 (CNPq).

6. Referências

Alencar, R.O. & Davis Jr, C.A., 2011. Geotagging aided by topic detection with Wikipedia. *14th AGILE Conference on Geographic Information Science*, Utrecht, The Netherlands, 461-478.

Amitay, E., Har'El, N., Sivan, R. & Soffer, A., 2004. Web-a-Where: Geotagging Web Content. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Sheffield, UK, 273-280.

Berners-Lee, T., 2006. *Linked Data - design issues* [online]. http://www.w3.org/DesignIssues/LinkedData.html [Accessed Aug 5, 2013].

Bizer, C., Heath, T. & Berners-Lee, T., 2009. Linked data - the story so far. *International Journal on Semantic Web and Information Systems,* 5 (3), 1-22.

Brin, S. & Page, L., 1998. The anatomy of a large hypertextual Web search engine. *Proceedings of the 7th International Conference on the World Wide Web*, Brisbane, Australia, 107-117.

Davis Jr, C.A., Pappa, G.L., de Oliveira, D.R.R. & de L Arcanjo, F., 2011. Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS,* 15 (6), 735-751.

Erwig, M., Güting, R. H., Schneider, M. & Michalis Vazirgiannis. 1999. Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases.*Geoinformatica* 3, 3 (September 1999), 269-296.

Goodchild, M. 2007c. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4): 211-221.

Jain, P., Hitzler, P., Yeh, P., Verma, K. & Sheth, A., 2010. Linked data is merely nore data. *AAAI Spring Symposium: linked data meets artificial intelligence*.

Keßler, C., Janowicz, K. & Bishr, M., 2009. An agenda for the next generation gazetteer: geographic information contribution and retrieval. *17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Seattle, Washington, 91-100.

Machado, I.M.R., Alencar, R.O., Campos Junior, R.O. & Davis Jr, C.A., 2010. An Ontological Gazetteer for Geographic Information Retrieval. *XI Brazilian Symposium on Geoinformatics*, Campos do Jordão (SP), Brazil, 21-32.

Machado, I.M.R., Alencar, R.O., Campos Junior, R.O. & Davis Jr, C.A., 2011. An ontological gazetteer and its application for place name disambiguation in text. *Journal of the Brazilian Computer Society,* 17 (4), 267-279.

McDougall, K. 2009. The potential of citizen volunteered spatial information for building SDI. GSDI 11, *Proceedings*..., Rotterdam, Netherlands.