QGISSPARQL - Integrando Dados Conectados e Sistemas de Informação Geográfica

Nerval de J. S. Junior[1] **, Sérgio Souza Costa**[2] **, Carlos Daniel dos S. Silva**[3]

[1] Universidade Federal do Maranhão – UFMA
Caixa Postal 322 – 65.086-110 – São Luıs – MA – Brasil

nerval@discente.ufma.br,sergio.costa@ufma.br,carlosdiel16@gmail.com

Abstract. *Geographic Information Systems handle a vast amount of data that could be available in connected data repositories. This paradigm enables the connection of data from various repositories, promoting reuse, and consequently, the replication of databases. This article aims to integrate Geographic Information Systems with connected data repositories. To achieve this, two plugins were developed and published on the official QGIS repository, enabling the publication of over 2.5 million triples in a public connected data repository (https://github.com/lambdageo).*

Resumo. *Sistemas de Informação Geográfica lidam com um grande volume de dados que poderiam estar disponíveis em repositórios de dados conectados. Esse paradigma permite a conexão de dados de diferentes repositórios, favorecendo o reuso e, por conseguinte, a replicação de bases de dados. Este artigo tem como objetivo a integração entre sistemas de informação geográfica e repositórios de dados conectados. Para isso, foram desenvolvidos e publicados dois plugins no repositório oficial do QGIS que permitiram a publicação de mais de 2,5 milhões de triplas em um repositório público de dados conectados (https://github.com/lambdageo).*

1. Introdução

O avanço da tecnologia da informação e a grande disponibilidade de computadores de grande potência facilitaram a análise de padrões espaciais e a visualização de dados geográficos. Atualmente, os denominados Sistemas de Informação Geográfica (SIG) realizam tratamentos computacionais e recuperação de dados geográficos, sendo importantes para diferentes tipos de usuários [Câmara et al. 2001, Carreiro Filho et al. 2022]. Nessa perspectiva, o artigo [Lopes et al. 2022], publicado no livro de minicursos da ERCEMAPI na edição de 2022, reforça o desenvolvimento de tecnologias para a análise de dados no espaço geográfico, destacando o aumento significativo na disponibilidade de dados com informações geoespaciais e compreendendo a necessidade de considerar o contexto espacial em que ocorrem os eventos geograficamente referenciados.

Exemplos de Sistema de Informação Geográfica proprietário destaca-se o ArcGis, conjunto integrado de softwares de SIG, produzido pela empresa americana ESRI (`https://www.esri.com/`), fornecendo ferramentas baseadas em padrões, para criação, gerenciamento, análise e visualização de dados geográficos

[Silva and MACHADO 2010]. Entre os gratuitos e de código aberto, pode-se citar iniciativas nacionais como o Sistema de Processamento de Informações Georreferenciadas (SPRING) [Câmara et al. 1996] e o Terralib/TerraView [Câmara et al. 2000], ambos desenvolvidos e mantidos pelo Instituto Nacional de pesquisas espaciais (INPE).

Atualmente, o QGIS tem se destacado na comunidade de software livre mediante diversos complementos. Ele permite a análise de dados georreferenciados, assim como a visualização e edição [Harumi-Ito et al. 2017]. Além de ser de código aberto, ele é altamente extensível por meio de plugins que adicionam funcionalidades extras ao software, como análise de terreno, geocodificação e integração com serviços de mapas. Este trabalho objetiva integrar sistemas de informação geográfica e repositórios de dados conectados por meio do desenvolvimento de plugins.

Com o uso dos Dados Conectados, é possível conectar dados de diferentes fontes e seguir padrões para representação dos dados que os tornam legíveis por máquinas, ou seja, os dados possuem semântica. Geralmente, os dados são dispostos em formato RDF (Resource Description Framework), o que permite interligar variados recursos usando vocabulários que aumentam a semântica desses relacionamentos [Isotani and Bittencourt 2015, Bandeira et al. 2015, Nascimento et al. 2020]. Esse paradigma cria inúmeras oportunidades para a integração semântica entre os próprios dados, motivando o desenvolvimento de novos tipos de aplicações e ferramentas, incluindo aplicações geográficas [Goodwin et al. 2008, Lopez-Pellicer et al. 2010, Kuhn et al. 2014, Garcia et al. 2019, Costa et al. 2016]. Um exemplo é a inclusão de suporte à consulta de dados espaciais através da especificação OGC-GeoSPARQL [Battle and Kolas 2011].

Ainda no contexto de aplicações geográficas, em [Garcia et al. 2019, Costa et al. 2016], os autores propuseram uma arquitetura denominada DBCells que utiliza o paradigma de dados conectados como uma abordagem para a publicação de dados de modelos de uso e cobertura da terra. De acordo com os autores, esta abordagem tem o potencial de aumentar o reuso e a replicabilidade de modelos que demandam um grande volume de dados. Contudo, a metodologia utilizada em [Garcia et al. 2019] requeria que os usuários escrevessem seus scripts de código para integrar o SIG com o repositório de dados conectados. Deste modo, espera-se que a integração proposta neste trabalho possa ser usada em diversos outros trabalhos que lidem com dados geográficos conectados.

2. Metodologia

Para a integração entre repositórios de dados conectados e sistemas de informação geográfica (SIG), foi escolhido o sistema QGIS. Por ser de código aberto e gratuito, além disso, é altamente extensível por meio de plugins que adicionam funcionalidades extras ao software, como análise de terreno, geocodificação e integração com serviços de mapas. Similarmente, este trabalho propõe o desenvolvimento de dois *plugins* utilizando a linguagem de programação Python, a biblioteca PyQt para criar interfaces interativas e a biblioteca rdflib[1] para trabalhar com dados RDF (Resource Description Framework). Após o desenvolvimento, estes plugins se integram à interface gráfica principal do QGIS, podendo ser acessados diretamente no menu principal, como mostra a Figura 1.

Do ponto de vista do usuário, os *plugins* desempenham duas tarefas principais:

[1]Mais informações em (`https://rdflib.readthedocs.io/`)

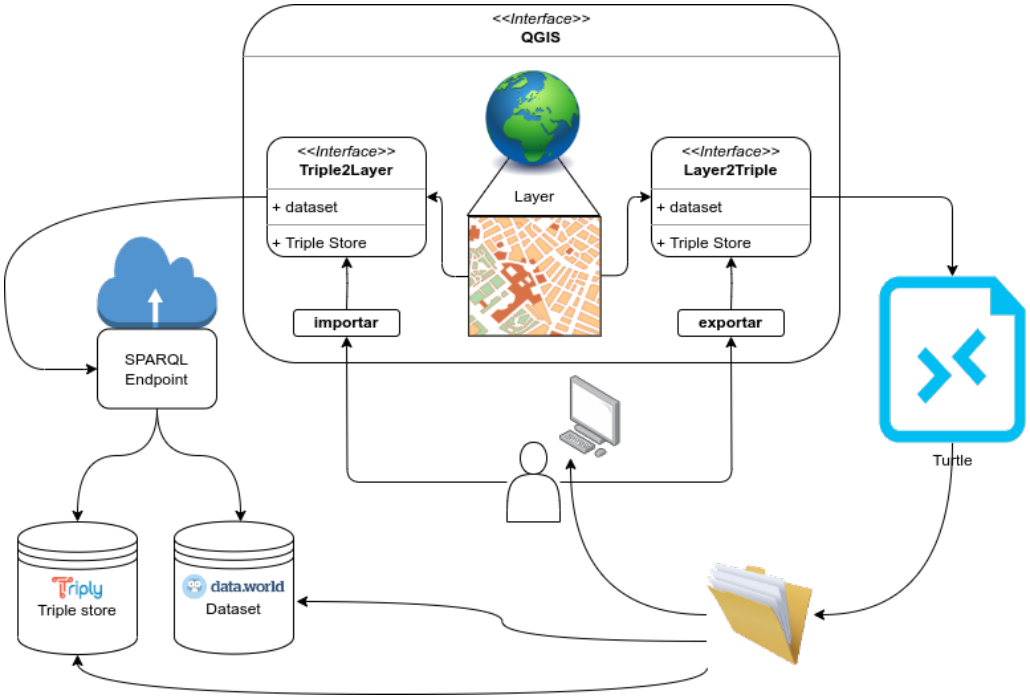

Figura 1. Arquitetura de *software*.

- **Importação**: o usuário irá fornecer algumas informações para que seja possível conectar a um repositório de dados conectados e, em seguida, transformar essas informações em uma camada geográfica no QGIS.
- **Exportação**: o usuário definirá como os dados de uma dada camada geográfica serão transformados em dados conectados, como o mapeamento dos atributos para termos de um dado vocabulário. Essa exportação será realizada por meio de arquivos, que poderão ser armazenados em uma pasta local ou enviados para repositórios de dados conectados.

A tarefa de importação é realizada pelo *plugin* Triple2Layer, enquanto a exportação é realizada pelo Layer2Triple. Nas próximas seções, serão detalhadas algumas decisões de arquitetura para cada um destes *plugins*.

2.1. QGISSPARQL:Triple2Layer

Este *plugin* visa importar dados conectados e convertê-los em camadas de dados geográficos no sistema de informações geográficas QGIS. Para o seu desenvolvimento, foram tomadas duas principais decisões de projeto. A primeira foi que ele seria compatível com servidores de bases de dados conectados, conhecidos como *triple stores* (ex: Virtuoso e Apache Jena *Fuseki*), e o portal de dados denominado *data.world* (`http://data.world`).

A segunda decisão foi sobre como os resultados de uma consulta seriam mapeados para dados em uma camada geográfica. Antes de descrever a solução, é importante compreender que uma base de dados conectados é composta por uma coleção de triplas utilizadas para estabelecer conexões entre conjuntos de dados geográficos, permitindo relacionar informações de forma semântica. Sua estrutura é composta por sujeito, predicado e objeto, como ilustrado na Figura 2.

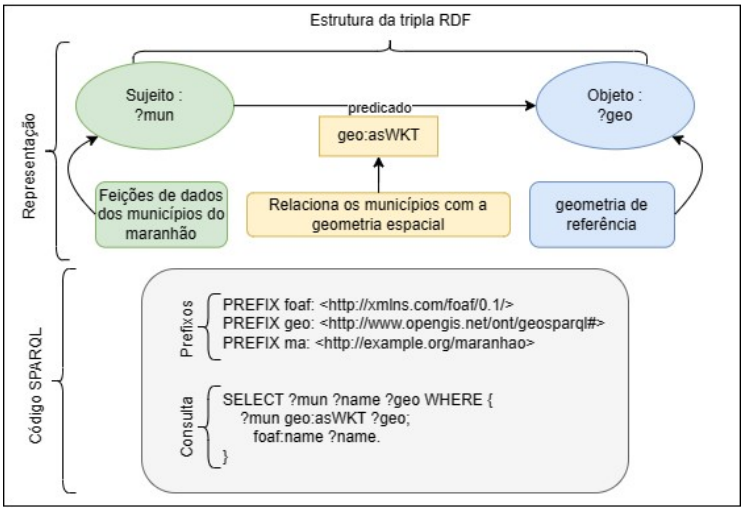

Figura 2. Ilustração da Tripla RDF e Consulta SPARQL.

O acesso a esses servidores de dados conectados é realizado através de uma linguagem de consulta denominada SPARQL (SPARQL Protocol and RDF Query Language)[2]. Por exemplo, a Figura 2 apresenta um código SPARQL com três variáveis `?mun`, `?name` e `?geo`. Essas variáveis poderiam ser mapeadas para: (1) geometrias, (2) identificador (chave primária da tabela) ou (3) atributo. O usuário poderá escolher mediante uma tabela de atributos (Figura 3) como cada variável da consulta será mapeada e, no caso de atributos, qual será o tipo de dados. Neste exemplo, a variável `?mun` será usada como chave primária, a variável `?name` será um atributo do tipo `string` e `?geo` possui os dados geométricos.

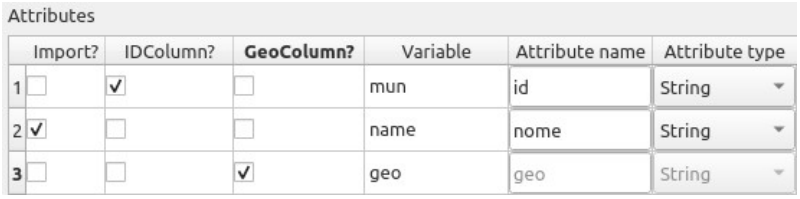

Figura 3. Mapeando as variáveis para camada geográfica

2.2. QGISSPARQL:Layer2Triple

Este *plugin* é responsável por gerar dados conectados a partir das informações contidas em dados vetoriais, incluindo pontos, linhas e polígonos. Não é possível exportar dados matriciais, como imagens de satélite. Os dados conectados são salvos em um arquivo local no formato Turtle[3], ou seja, não serão enviados diretamente para um repositório de dados conectado. Ele possui um nível de complexidade maior em comparação com o Triple2Layer, e foram tomadas as seguintes decisões de projeto.

[2]Ela é uma linguagem de consulta padronizada pelo World Wide Web Consortium (W3C) para recuperar, consultar e manipular dados armazenados em formato RDF (Resource Description Framework) na Web Semântica.

[3]É uma representação textual de dados RDF (Resource Description Framework) com uma sintaxe mais legível para humanos.

2.2.1. Carregamento dos vocabulários

Os usuários poderão carregar vocabulários específicos, podendo incluir aqueles criados pelo próprio usuário. Logo, a ferramenta precisa possibilitar o carregamento destes vocabulários, inicialmente, nos dois formatos de serialização mais comuns: Turtle e XML. Na Figura 4 é mostrado um exemplo, onde está sendo carregado o vocabulário OGC-GeoSPARQL.

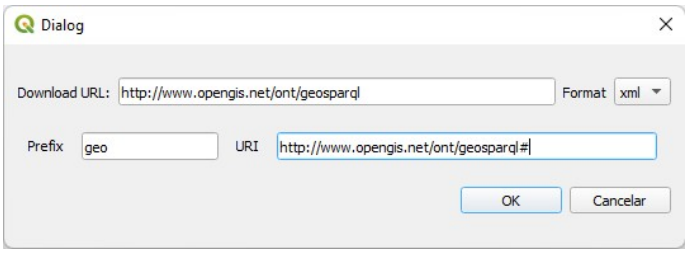

Figura 4. Caixa de diálogo para carregamento de vocabulários

Ainda não será possível o envio de arquivos locais, contudo essa funcionalidade poderá ser incluída em versões futuras.

2.2.2. Definindo a classe e um identificador único para as feições

Uma camada geográfica de dados vetoriais é composta por feições geográficas (*features*), as quais serão mapeadas como recursos pertencentes a uma dada classe e que possuem uma URI (identificador único). Portanto, será necessário permitir ao usuário definir a classe e a estratégia para a geração das URIs. No caso da classe, o usuário poderá selecionar entre todas as classes dos vocabulários carregados previamente, e para a geração da URI, foram consideradas duas estratégias. Em ambas as estratégias, o usuário irá definir uma URL base que será concatenada com uma das seguintes informações:

- **Universally unique identifier (UUID)** é um identificador único universal que é utilizado para identificar informações, objetos ou entidades de forma globalmente única. Neste caso, será gerado um UUID para cada feição.
- **Atributo da camada geográfica**, geralmente já existe um atributo usado como identificador para a camada, que poderá ser selecionado.

Na Figura 5 os elementos da interface gráfica relacionados à escolha da classe e à estratégia de geração do identificador único.

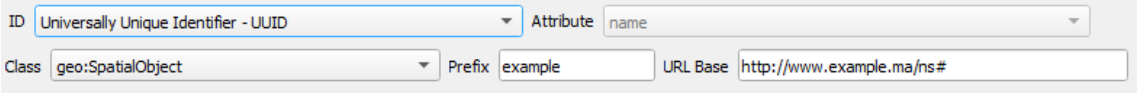

Figura 5. Selecionando a classe e a estratégia de geração das URIs

Usando a estratégia baseada em UUID, como na Figura 5, poderá ser gerada uma URI como: `http://example.ma/ns#02e64008-877f-489c-b0a3-2bd592f91698`. A serialização em Turtle usará prefixos, como no Código 1.

```
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix example: <http://example.ma/ns#> .

example:02e64008-877f-489c-b0a3-2bd592f91698 a geo:SpatialObject ;

```

Código 1: Exemplo de uma feição usando a estratégia de UUID

2.2.3. Exportando geometrias

Na interface gráfica haverá uma opção que permitirá ao usuário escolher se deseja exportar as geometrias de cada feição. Neste caso, os dados serão exportados utilizando o formato de serialização WKT (Well-Known Text). Este é um formato de representação textual utilizado para descrever geometrias espaciais em Sistemas de Informação Geográfica (SIG) e sistemas de georreferenciamento. Ele fornece uma maneira padronizada e legível por humanos de representar pontos, linhas, polígonos e outros tipos de geometrias em um espaço bidimensional ou tridimensional.

Cada geometria em uma camada geográfica será relacionada com sua representação no formato WKT através da propriedade `http://www.opengis.net/ont/geosparql#asWKT`, que pertence o vocabulário OGC-GeoSPARQL [Perry and Herring 2011]. Nessa versão não será possível escolher o formato e nem a propriedade RDF.

2.2.4. Conectando as propriedades

Além das geometrias, um dado vetorial está associados a outros atributos, como mostra a Figura 6. Neste exemplo, cada município é associado a um código, nome, sigla da unidade federativa e área.

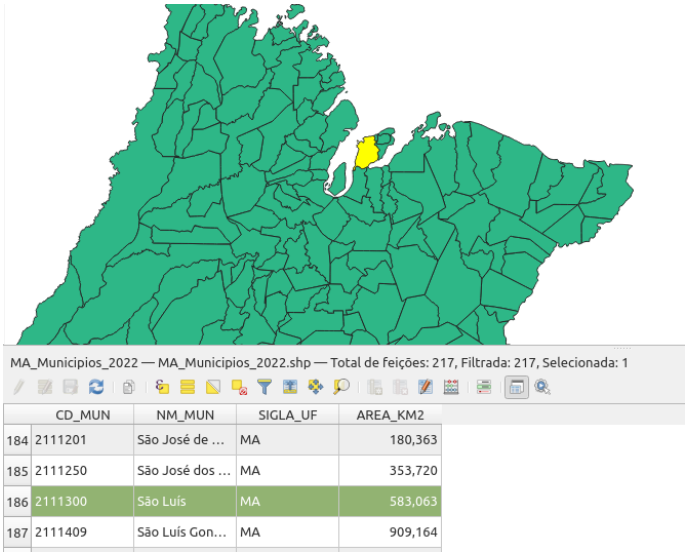

Figura 6. Divisão municipal do estado do Maranhão

Estes atributos podem ser associados a propriedades RDF dos vocabulários carre-

gados previamente. De modo geral, consideram-se três estratégias:

1. **Atributo da camada**: a propriedade será mapeada para um valor literal presente na tabela de atributos da camada geográfica;
2. **Valor Constante**: todas as feições (no exemplo do estado, seriam todos os municípios) serão mapeadas para o mesmo valor que será inserido em uma caixa de texto;
3. **Vocabulário**: todas as feições serão associadas a alguma classe de um dos vocabulários carregados.

A Figura 7 mostra um exemplo de cada uma das estratégias: (1) conecta a propriedade `lgd:officialName` para um atributo da camada geográfica NM_MUN (nome do munícipio), (2) conecta a propriedade `lgd:country` para uma literal, no caso, a palavra "Brasil" e por fim (3) conecta uma propriedade a uma classe, no caso `lgd:map_type` para `lgd:Municipality`.

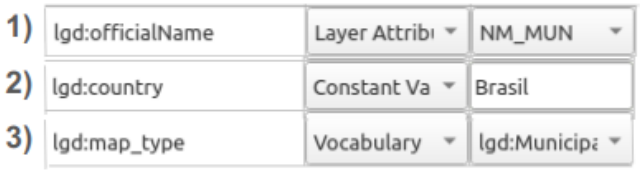

Figura 7. Exemplos das estratégias de conexão

3. Resultados

Ambos os plugins já estão disponíveis no repositório do QGIS, tanto o Triple2Layer, quanto o Layer2triple. Além disso, ambos estão publicados como software de código aberto em `https://github.com/LambdaGeo/`.

3.1. Publicando dados geográficos como dados conectados

Estes plugins estão sendo utilizados pelo projeto DBCells [Costa et al. 2017] e foram utilizados por [Silva 2023] para exportar os dados de cobertura da terra de todo o país, que foram utilizados em um modelo publicado por [Silva Bezerra et al. 2022]. Além dos dados de cobertura, foram exportadas as variáveis de distância aos rios e distância às sedes urbanas.

Os dados de cobertura dos diferentes anos estão sendo agrupados em `https://data.world/lambdageo/luccmebrlanduse/`, com quase 3 milhões de triplas. As outras variáveis explicativas utilizadas pelo modelo estão sendo agrupadas e publicadas em `https://data.world/lambdageo/luccmebrdrivers`, que atualmente são apenas duas.

3.2. Carregando dados conectados como camadas geográficas

Uma das principais vantagens do paradigma de dados conectados é a possibilidade de um dado fazer referência a dados de outras coleções. A proposta do projeto DBCells é que os dados dos modelos possam ficar em coleções diferentes daquelas onde o DBCells estará publicado. A proposta do projeto é que os dados das células, incluindo sua componente espacial, estejam disponíveis em um servidor dedicado. Atualmente,

estão sendo testados alguns protótipos com o objetivo de validar este trabalho, e a versão testada está sendo parcialmente disponibilizada através de uma conta gratuita em: `https://dbcells-fuseki-production.up.railway.app/`.

Como destacado na Seção 2, o plugin Triple2Layer é utilizado para carregar os dados de um ou mais repositórios de dados conectados através de consultas SPARQL. Para conectar dados de mais de um repositório, é utilizada a palavra reservada SERVICE. O Código 2 apresenta um exemplo de uma consulta que integra Data World e DBCells.

```
SELECT ?cell ?agric ?wkt
where {
    SERVICE <https://dbcells-fuseki-production.up.railway.app/cells>
    ↪ {
        ?cell geo:asWKT ?wkt.     }
    ?s1 a qb:Observation;
            dbc-measure:mean ?agric;
            qb:dataSet ?ds1;
            sdmx-dimension:refArea ?cell.
    ?ds1 dbc-attribute:feature dbc-code:landcover-agric.
}
```

Código 2: Integrando os dados do DataWorld e DBCells

A consulta do Código 2 irá carregar os dados de geometria do DBCells e a porcentagem de áreas de agricultura do DataWorld. A Figura 8 apresenta o dado importado e com uma legenda indicando o percentual de agricultura em cada célula.

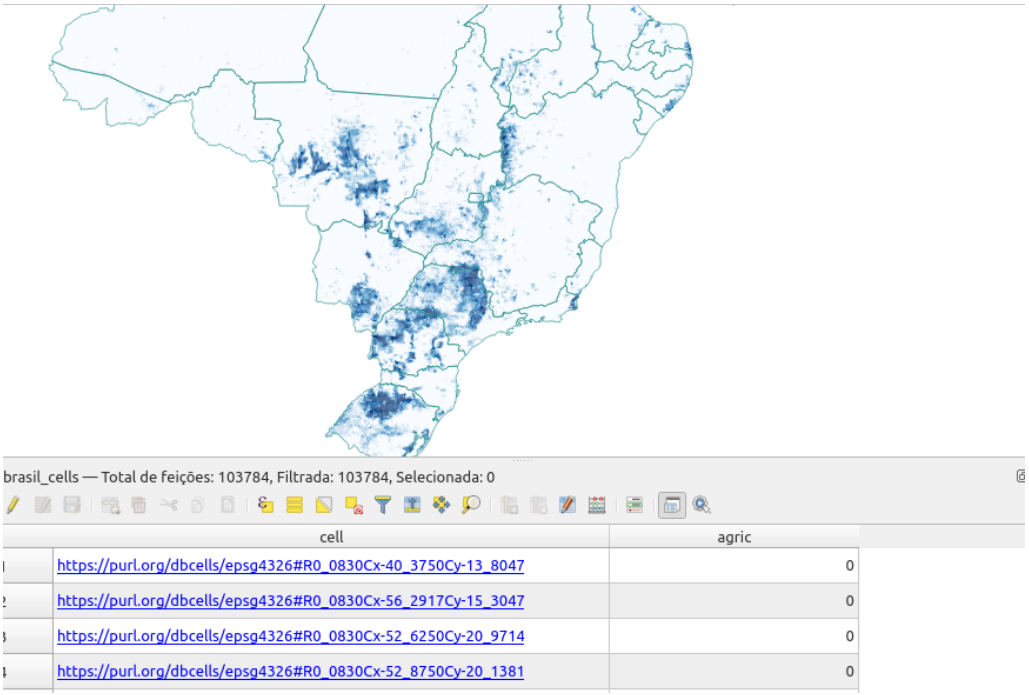

Figura 8. Camada geográfica carregada a partir de um repositório de dados conectados

4. Conclusão

Este trabalho apresentou uma integração entre sistemas de informação geográfica e repositório de dados conectados. Essa integração foi realizada através da criação de dois plugins desenvolvidos para o software QGIS: Triple2Layer e Layer2Triple. Os plugins permitiram, respectivamente, importar dados de um repositório de dados conectados e exportar uma camada geográfica como dados conectados. Ambos já estão publicados no repositório oficial do QGIS, e podem ser instalados diretamente da interface do sistema.

O Triple2Layer permite importar dados de servidores *triple store* e/ou do portal de dados `Data.World` a partir de consultas SPARQL. Através de sua interface, o usuário define como o resultado da consulta será convertido para a camada geográfica. O Layer2Triple permite o carregamento de vocabulários que serão usados para a exportação da camada geográfica para dados conectados. Na versão 0.1 os dados são exportados no formato Turtle e as geometrias são serializadas em WKT.

Esta abordagem de integração está sendo utilizada em um projeto denominado DBCells, e neste artigo foram apresentados alguns dados geográficos publicados como dados conectados através da abordagem aqui proposta. Os dados são de um modelo de uso e cobertura da terra publicado em [Silva Bezerra et al. 2022]. No portal já foram publicados mais de 2 milhões de triplas que podem ser acessados em `https://data.world/lambdageo/`. Estes dados publicados podem ser carregados novamente para o sistema de informação geográfica através do Triple2Layer demonstrando assim a integração completa entre estes sistemas e repositórios de dados conectados.

Agradecimentos

Este estudo foi viabilizado por meio do apoio financeiro provido pelas bolsas de iniciação científica e tecnológica concedidas pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) e pela Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (Fapema).

Referências

Bandeira, J. M., Alcantara, W., Barbosa Sobrinho, A., Ávila, T. J. T., Bittencourt, I., and Isotani, S. (2015). Dados abertos conectados. *III Simpósio Brasileiro de Tecnologia da Informação*.

Battle, R. and Kolas, D. (2011). Geosparql: enabling a geospatial semantic web. *Semantic Web Journal*, 3(4):355–370.

Câmara, G., Davis, C., Monteiro, A. M. V., and D'ALGE, J. (2001). Introdução à ciência da geoinformação. *São José dos Campos: INPE*, 345.

Câmara, G., Souza, R., Pedrosa, B., Vinhas, L., Monteiro, A. M. V., Paiva, J., Carvalho, M. d., and Gattass, M. (2000). Terralib: Technology in support of gis innovation. In *II Workshop Brasileiro de Geoinformática, GeoInfo2000*, volume 2, pages 1–8. São Paulo.

Câmara, G., Souza, R. C. M., Freitas, U. M., and Garrido, J. (1996). Spring: Integrating remote sensing and gis by object-oriented data modelling. *Computers Graphics*, 20(3):395–403.

Carreiro Filho, F. B., Borges, H. P., and Cortes, O. A. C. (2022). Paralelizacao eficiente na simulacao da elevacao do nível do mar em areas de reentrancias maranhenses. In *Anais da X Escola Regional de Computação do Ceará, Maranhão e Piauí*, pages 99–108. SBC.

Costa, S. S., Moreira, E. G., da Silva, M. L., de Sousa Lima, T. M., and Luıs-MA-Brazil, S. (2016). Dbcells-an open and global multi-scale linked cells. In *GeoInfo*, pages 1–11.

Costa, S. S., Silva, M. L., Lima, T. M. d. S., and Moreira, E. G. (2017). Dbcells - an open and global multi-scale linked cells. *Revista Brasileira de Cartografia*, 69(5).

Garcia, D. A., Costa, S. S., and Moreira, E. G. (2019). Publicação de dados conectados para modelos de uso e cobertura da terra/linked data publication for land change models. *REVISTA GEONORTE*, 10(36):77–94.

Goodwin, J., Dolbear, C., and Hart, G. (2008). Geographical linked data: The administrative geography of great britain on the semantic web. *Transactions in GIS*, 12:19–30.

Harumi-Ito, M., Filho, H. F., and Conti, L. A. (2017). Uso do software livre qgis (quantum gis) para ensino de geoprocessamento em nível superior. page 127–148.

Isotani, S. and Bittencourt, I. I. (2015). *Dados abertos conectados: em busca da web do conhecimento*. Novatec Editora.

Kuhn, W., Kauppinen, T., and Janowicz, K. (2014). Linked data-a paradigm shift for geographic information science. In *Geographic Information Science: 8th International Conference, GIScience 2014, Vienna, Austria, September 24-26, 2014. Proceedings 8*, pages 173–186. Springer.

Lopes, G. R., Pelarigo, K. J., Delbem, A. C., and de Sousa, J. B. (2022). Análise exploratória de dados espaciais com python. *Sociedade Brasileira de Computação*.

Lopez-Pellicer, F. J., Silva, M. J., Chaves, M., Javier Zarazaga-Soria, F., and Muro-Medrano, P. R. (2010). Geo linked data. In *Database and Expert Systems Applications: 21st International Conference, DEXA 2010, Bilbao, Spain, August 30-September 3, 2010, Proceedings, Part I 21*, pages 495–502. Springer.

Nascimento, L., Castro, P., Oliveira, M., Jose, F., Costa, V., Moura, C., Freitas, R., and Monteiro, O. (2020). Pixel, plataforma para integração de experimentos de interoperabilidade em sistemas legados de saúde pública. In *Anais da VIII Escola Regional de Computação do Ceará, Maranhão e Piauí*, pages 181–188. SBC.

Perry, M. and Herring, J. (2011). Ogc geosparql-a geographic query language for rdf data, ogc implementation standard.

Silva, C. D. d. S. (2023). Um vocabulário para modelos de uso e cobertura baseado no data cube vocabulary.

Silva, V. C. B. and MACHADO, P. d. S. (2010). Iniciando no arcgis. *Belo Horizonte: Centro Universitário de Belo Horizonte*.

Silva Bezerra, F. G., Von Randow, C., Assis, T. O., Bezerra, K. R. A., Tejada, G., Castro, A. A., Gomes, D. M. d. P., Avancini, R., and Aguiar, A. P. (2022). New land-use change scenarios for brazil: Refining global ssps with a regional spatially-explicit allocation model. *PLOS ONE*, 17(4):1–17.