Um *Dataset* Enriquecido com Dados Extraídos da *Web* para Aplicações de Georreferenciamento

Clovis S. Junior[1], **Carina F. Dorneles**[2]

[1]Instituto de Ciências Exatas e Naturais/UFR
Rondonópolis, MT/Brasil

[2]Departamento de Informática e Estatística - INE
Universidade Federal de Santa Catarina - UFSC/Florianópolis

clovis@ufr.edu.br, carina.dorneles@ufsc.br

Abstract. *Agricultural and environmental applications largely depend on georeferenced data. Getting this type of data requires high resources related to hardware and specialized human resources. Data extraction can be a viable alternative for creating datasets for this demand. It is possible to find public repositories on the Web to create or complement datasets in the agricultural and environmental domain, whether for delimiting agricultural areas or identifying and monitoring environmental areas. This paper presents a proposal for data extraction from the Web to create a dataset for agricultural and environmental use through geo-coordinates extraction in public repositories.*

Resumo. *Aplicações agrícolas e ambientais dependem de dados georreferenciados. A obtenção desse tipo de dado exige recursos elevados relacionados a hardware e recursos humanos especializados. A extração de dados da Web pode ser uma alternativa viável para criação de datasets para essa demanda. É possível encontrar repositórios públicos em ambiente Web para criar ou complementar datasets no domínio agrícola e ambiental, seja para delimitação de áreas agrícolas ou identificação e monitoramento de áreas ambientais. O presente artigo apresenta uma proposta para extração de dados da Web com o objetivo de criar um dataset para uso agrícola e ambiental por meio da extração de geo-coordenadas em repositórios públicos.*

1. Introdução

O mapeamento de áreas agrícolas e ambientais usualmente é realizado com visitas aos locais investigados. A identificação física das áreas de interesse consiste em determinar geograficamente os respectivos limites. Outro ponto importante refere-se a coleta de dados georreferenciados que demandam custos altos sendo impeditivo para um grande número de propriedades rurais. Essa dificuldade resulta em irregularidades compulsórias de pequenas propriedades, em geral associadas à agricultura familiar, impossibilitando repasses de recursos assistenciais em diferentes esferas governamentais. A delimitação dessas áreas, em alguns casos, pode ser encontrada em repositórios web, a exemplo desses pode-se citar:

- AIDDATA: Tem a missão de conectar tomadores de decisão e pesquisadores que compartilham interesses em trabalhar juntos usando dados granulares e ferramentas inovadoras para resolver problemas, direcionar recursos com precisão e usar

evidências para medir impactos pretendidos ou não intencionais relacionados a área econômica, site: `https://www.aiddata.org/datasets`

- GeocodedData: O propósito desse grupo de trabalho é cria impacto em comunidades do continente Africano, informando iniciativas governamentais, índices globais, manchetes e agendas de desenvolvimento. Também preparam outros grupos para trabalhar em ações de desenvolvimento ancorados às realidades relacionadas com a realidade africana `https://www.afrobarometer.org/data/data-sets/`.

Repositórios públicos possibilitam a criação ou a complementação de *datasets* nesse domínio. O primeiro passo para realizar a extração consiste em identificar repositórios de interesse referente ao domínio investigado. Nesse contexto, o principal problema refere-se a localização de dados georreferenciados na Web, porque em geral são dados de acesso restrito.

Fontes de dados são disponibilizados na web de forma pública como tabelas com dados geográficos para uma variedade de domínios, como sustentabilidade urbana, redes de transporte, estudos políticos e saúde [Cruz et al. 2013]. Neste artigo, é apresentada uma alternativa para criar um *dataset* com dados georreferenciados, contribuindo com a identificação e monitoramento agrícola e ambiental. Nesse contexto, os dados georreferenciados foram obtidos por meio de extração de dados da Web em portais governamentais de acesso público sem custos ou licenças comerciais.

A criação do *dataset* proposto seguiu alguns critérios, como a abordagem feita por [Azad et al. 2018], que indica alguns processos de enriquecimento úteis para essa demanda, sendo:

- Fusão de dados: unificação de dados de várias fontes que representam a mesma entidade com consistência e representação útil.
- Reconhecimento da entidade de dados: processo de identificação de palavras em textos. Encontra e reconhece nomes de pessoas, empresas, organizações, cidades e outros tipos predefinidos de entidades.
- Desambiguação de dados: desambiguação ou eliminação de duplo sentido é o processo de identificação da entidade correta de dados para variações não uniformes e ambíguas em nomes de entidades.
- Segmentação de dados: processo de agrupar dados de acordo com um conjunto de atributos predefinidos.
- Imputação de dados: estima valores para itens de dados ausentes ou conflitantes.
- Categorização de dados: identifica dados em diferentes categorias com base em tópicos, eventos ou outras características.

O artigo também utiliza reconhecimento da entidades ou *tags* de dados para identificação de coordenadas úteis para o *dataset* proposto, conforme apresentado na Figura 1. Embora pesquisas abordem a extração de dados *Web* para problemas diferentes as dificuldades são semelhantes, como identificar dados relevantes para o domínio investigado [Dong et al. 2020]. O artigo propõe uma alternativa para a criação de um *dataset* a partir do enriquecimento com arquivos de dados georreferenciados. O público alvo para os *dataset* são técnicos agrícolas,engenheiros florestais e profissionais envolvidos com monitoramento e perícia ambiental, a proposta é detalhada na seção 3.

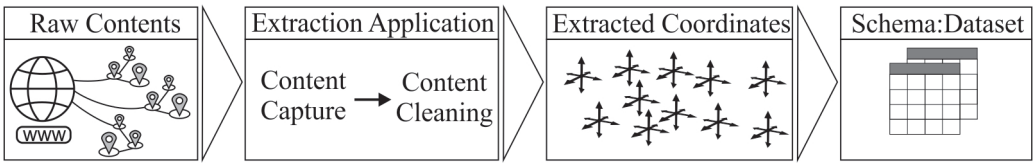

Figura 1. Visão Geral da Proposta.

Este artigo está organizado como segue. Na Seção 2 são descritos os trabalhos relacionados com pesquisas relacionadas com extração de dados. Na Seção 3, são apresentadas as estruturas dos dados brutos e proposta para a criação do *dataset* com os métodos utilizados. Os dados gerados para o *dataset* e exemplos de aplicações para uso dos mesmos são apresentados na Seção 4. Na Seção 5, são apresentadas as conclusões e propostos alguns trabalhos futuros.

2. Trabalhos relacionados

Esta seção apresenta uma revisão sobre abordagens para criação de *datasets* com extração de dados de conteúdos web utilizando a estrutura *XML* como referência para interpretação de conteúdos em *Keyhole Markup Language* ou *KML*.

O tema central do artigo está associado à criação de datasets utilizando extração de dados web com dados georreferenciados. O recurso de extrair dados não é algo recente, segundo [Lloret-Gazo 2020] a extração de dados de páginas Web teve início nos anos 2000, e as soluções propostas atualmente tem sido essencialmente o uso de *wrappers* implementados como programas ou bibliotecas para extração de dados. Ainda, segundo o autor existem muitos trabalhos publicados sobre extração de dados de documentos Web e várias pesquisas sobre esse assunto.

Segundo [Azeroual and Jha 2021], o grande problema na extração reside na pouca qualidade dos dados, isso é derivado por diferentes razões e representa um desafio que não deve ser subestimado. Em certos casos, os dados estão no formato errado ou no intervalo de valores errado, para resolver isso os dados devem ser limpos ou validados. A alta qualidade dos dados não é algo apenas desejável, mas um dos principais critérios para determinar se o enriquecimento foi bem sucedido e as afirmações obtidas estão corretas. Nesse contexto entende-se por enriquecimento a associação dos dados à entidades que armazenam dados de propriedades rurais, resultando na agregação de recursos para criação de informações ambientais. Complementarmente, [Jaya et al. 2017] indica que alta qualidade de dados é obtida quando dados são adequados para uso e capaz de atender ao objetivo definido por usuários de dados. Esta definição sugere claramente que a qualidade dos dados é altamente dependente do contexto, sinergia, necessidades, capacidades de uso e acesso.

Um ponto importante refere-se à escolha do tipo de dado para utilizar. Nesse sentido, optou-se por utilizar dados georreferenciados motivado pelas características do domínio investigado com interesse agrícola e ambiental. O artigo apresentado por [Gong et al. 2017] faz uma abordagem semelhante, entretanto utiliza dados semi-estruturados de forma mais genérica, a abordagem dessa proposta é utilizar dados estruturados de forma específica para o meio ambiente.

O artigo [Lloret-Gazo 2020] apresenta uma solução independente de modificações

na estrutura da fonte de dados, isto porque os *wrappers* tradicionais dependem muito de estruturas como *HTML* como fontes, para isso, são utilizadas regras sintáticas para identificação dos conteúdo de interesse. Outras iniciativas para o uso de geo-coordenadas são descritas em [Imbrenda et al. 2013], no qual apresenta um software de código aberto denominado "Free and Open Source Software for land degradation vulnerability assessment (FOSS)", para identificar diferentes níveis de vulnerabilidade com modelos de áreas sensíveis ao meio Ambiente (ESAs). Outra iniciativa é apresentada por [openforis 2021], com o "Open Foris", sendo este um conjunto de ferramentas de software gratuitas de código aberto que facilita a coleta, análise e relatório de dados sobre inventários florestais.

As informações apresentadas no final da seção 4 referem-se a verificações regionais de localizações baseadas em dados enriquecidos com extração web e dados importados de *datasets* governamentais, como exemplo pode-se citar:

- Informações estatísticas sobre o plantio, colheita, produção e rendimento médio, de forma sistemática, para os principais produtos das lavouras permanentes e temporárias. Dados provenientes de pesquisa de previsão e acompanhamento de áreas, produção e rendimento médio de 25 importantes produtos agrícolas, desde a fase de intenção de plantio até o final da colheita, `https://dados.gov.br/dataset/la-levantamento-sistematico-da-producao-agricola-lspa`
- SIPAF, identifica os produtos oriundos da agricultura familiar, que vem crescendo e se organizando para produzir cada vez mais e com mais qualidade, `https://dados.gov.br/dataset/sipaf-selo-de-agricultura-familiar`

Nesse cenário a aplicação ilustrativa limita-se a verificar distanciamentos entre pontos. Destaca-se que a aplicação proposta no artigo trata de extração, enriquecimento e operações com georreferenciamento de regiões em áreas ambientais, as demais iniciativas citadas tem foco em análise de solos e dados referentes a inventários, dessa forma a abordagem do artigo tem contribuição em áreas diferentes.

3. Arquitetura para Criação do *Dataset* Georreferenciado Proposto

A proposta apresenta uma solução para criação de um *dataset* com dados georreferenciados de fontes externas, favorecendo a criação de informações nos domínios agrícola e ambiental associados às necessidade de georreferenciamento. Outro ponto importante refere-se a possibilidade de contribuir com informações georreferenciadas à demandas governamentais como delimitação de área de preservação permanente, nascentes de rios, sedes de propriedades e delimitação de bordas de propriedades. Essa demanda foi normatizada no Código Florestal Brasileiro (Lei no 12.651) alterações na delimitação das Áreas de Preservação Permanente ou APPs para imoveis rurais criando novos conceitos como área rural consolidada que também são aplicados às APP parametrizados pelo modulo fiscal de cada município. Nessa normatização foram criadas cinco classes de modulo fiscal no qual os dispositivos legais a serem aplicados nas nascentes, cursos e corpos d´água ficam mais restritivos [Bonamigo 2015].

3.1. Metodologia

O desenvolvimento do trabalho teve início com a identificação do tipo de fonte de dados adequada para o domínio abordado. Assim, optou-se por fontes da *Web* com conteúdo relacionado a georreferenciamento para áreas rurais utilizando arquivos com formato *keyhole markup language* ou *KML*, sendo este baseado na estrutura *eXtensible Markup Language* ou *XML*. A escolha do formato *XML* é justificada pela disponibilidade de dados para essa demanda específica, outros formatos de arquivos são igualmente importantes e acessíveis pelas principais ferramentas de GIS, a Figura 2 apresenta a disponibilidade de formatos encontrado em um repositório governamental[1] utilizado como complemento para a obtenção de alguns dados utilizados neste artigo. Posteriormente, foi desenvolvido um protótipo para extração de dados das fontes escolhidas. A etapa final consistiu em utilizar os dados extraídos para criar o *dataset* para uso em aplicações no domínio agrícola.

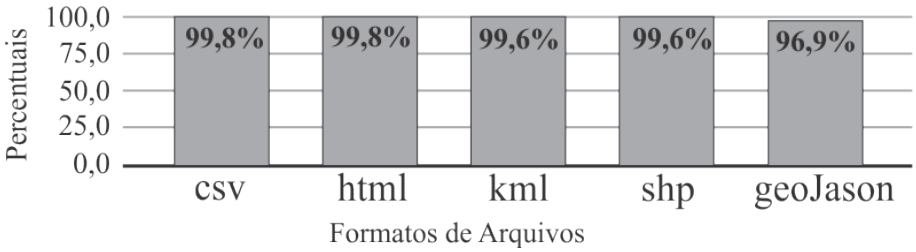

Figura 2. Disponibilidade de formatos de arquivos.

O extrator utilizado no artigo é uma ferramenta complementar desenvolvida especificamente para essa demanda (o código fonte pode ser acessado no repositório `https://github.com/clovissjunior/Geo`) com o objetivo de preparar adequadamente os dados para armazená-los na estrutura de banco de dados de destino. Esses dados poderão ser usados para criação de relatórios com georreferenciamento relacionado às áreas degradadas, áreas de preservação permanente, delimitação de áreas de plantio entre outros. A Figura 3 mostra a interface desenvolvida para extração, limpeza dos dados georreferenciados para criação do *dataset*.

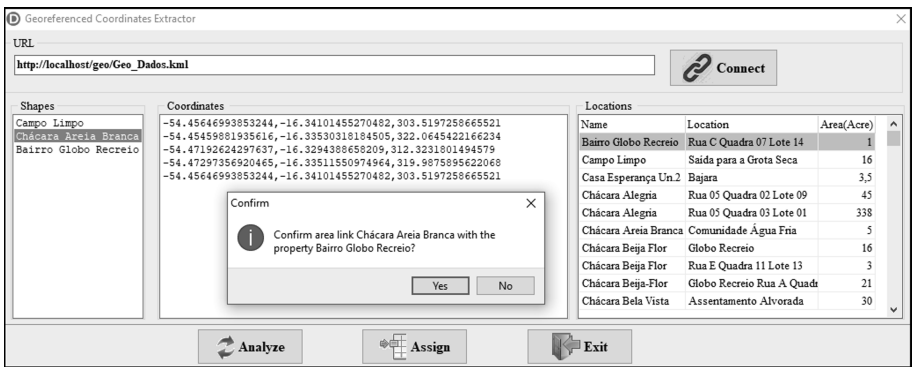

Figura 3. Interface da Aplicação de Extração de Dados.

[1]https://dados.gov.br/dataset?tags=Geoespacial

3.2. Infraestrutura dos Dados

A arquitetura da aplicação apresentada no artigo faz a extração dos dados em três etapas:

- A primeira obtém dados de fontes externas a partir de um endereço (*url*) indicado copiando o conteúdo para uma estrutura local.
- A segunda etapa refere-se a limpeza dos dados, após o conteúdo ser capturado é realizada uma análise e os dados desnecessários são removidos utilizando um micro parse para identificar os conteúdos relevantes como nomes de áreas e coordenadas geográficas.
- A etapa final refere-se à criação da estrutura física para o armazenamento dos dados georreferenciados no esquema de destino.

O armazenamento dos dados é feito com intervenção do usuário, para indicar a correspondência entre o valor dos atributos da entidade e as coordenadas georreferenciadas capturadas. A Seção 4 apresenta o uso das coordenadas no cenário ambiental. A Figura 4 mostra a estrutura dos dados após a extração sem a limpeza dos mesmos.

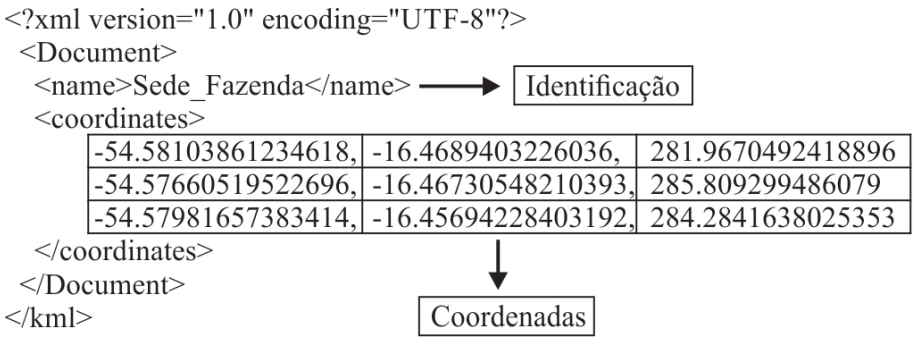

Figura 4. Fonte de Dados Original.

Optou-se por criar uma arquitetura de software para extração, limpeza e associação dos dados georreferenciados por motivo das características específicas do cenário investigado. O desenvolvimento próprio possibilita a criação de uma solução adequada às demandas que se quer explorar. O software também realiza pesquisas para calcular a distância de pontos importantes entre áreas degradadas e áreas urbanas ou de relevância ambiental.

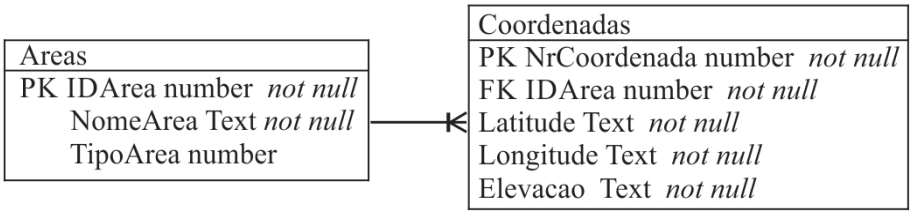

Figura 5. Estrutura para o Dataset.

4. Resultados

O artigo apresenta um alternativa para criação de um *dataset* com dados georreferenciados utilizando extração de dados como recurso. Conforme descrito na Seção 1, os dados podem favorecer aplicações para o meio ambiente e agricultura, auxiliando a identificação de áreas de interesse como reservas de preservação ambiental, áreas de plantio e áreas degradadas.

4.1. Extração dos Conteúdos

Optou-se pela extração de dados Web em razão da relevância para o contexto no qual a pesquisa se desenvolveu, sendo esta associada a aplicações e soluções agrícolas e ambientais na região centro-oeste do Brasil. A Figura 6 mostra os dados capturados em um estágio inicial sem qualquer transformação. Nesta fase, os dados ainda não tem utilidade, pois estão dispostos na estrutura original sem qualquer limpeza ou modificação para torná-los acessíveis por aplicativos de terceiros.

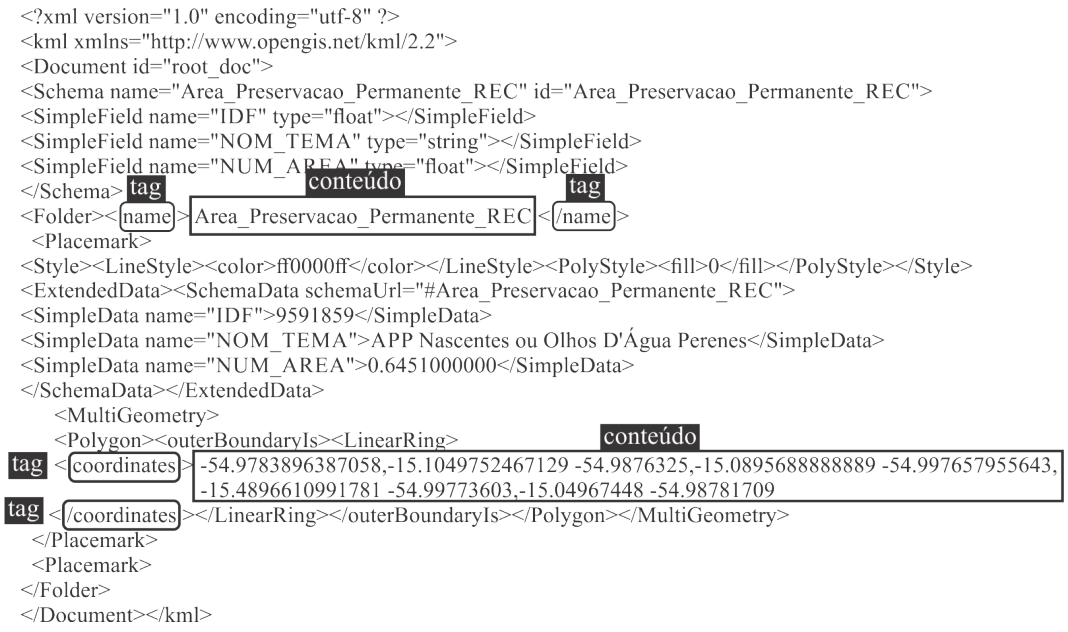

Figura 6. Conteúdo Original dos Arquivos de Dados.

4.2. Limpeza dos Dados

Conforme apresentado na Seção 3 os dados utilizados foram extraídos de arquivos *KML*, e os testes foram realizados em um servidor Web local. Destaca-se que esse formato de arquivo é utilizado e compartilhado por diversos aplicativos utilizados para geoprocessamento. A estrutura *KML* contempla diversos dados ou *tags*, apesar disso, foram utilizados apenas dois, conforme mostra a Figura 6 e também a Figura 7, sendo estes o nome da área ou ponto georreferenciado e as coordenadas que delimitam a mesma. As coordenadas são os dados mais importantes para a extração realizadas junto à base de dados para a demonstração dos resultados. A plicação desenvolvida para realizar a limpeza dos dados realiza uma pesquisa no arquivo de dados para identificar as *tags* com nomes e coordenadas, somente esses dois dados são utilizados para a criação do *dataset*.

https://Uniform_Resource_Locator/data.kml

Latitude	Longitude		Latitude	Longitude	
-55.70101499034053	-16.89472397864652	Area 1	-55.67952081209619	-16.91435778517345	Area 2
-55.69083781135536	-16.91544069960135		-55.67829453310166	-16.91322983329321	
-55.68711889758733	-16.91598712555625		-55.67751198398959	-16.91269029991324	
-55.68243207960381	-16.91523801131084		-55.67624808991618	-16.91176189376817	

Figura 7. Limpeza de Conteúdo.

4.3. Armazenamento Temporário de Dados

Após a extração e limpeza dos lados optou-se por armazená-los temporariamente em um banco de dados local. A justificativa para o uso desse banco de dados intermediário é a possibilidade do armazenamento estruturado dos dados. Na fase de limpeza os dados são armazenados temporariamente em arquivos textuais. O uso de um banco de dados temporário permite a simplificação na associação entre as entidades da base de dados para enriquecimento e os dados locais, pois ambos são dados estruturados em entidades e atributos, conforme apresentado na Figura 8. Em outra etapa da extração, os dados temporários são eliminados para novas operações.

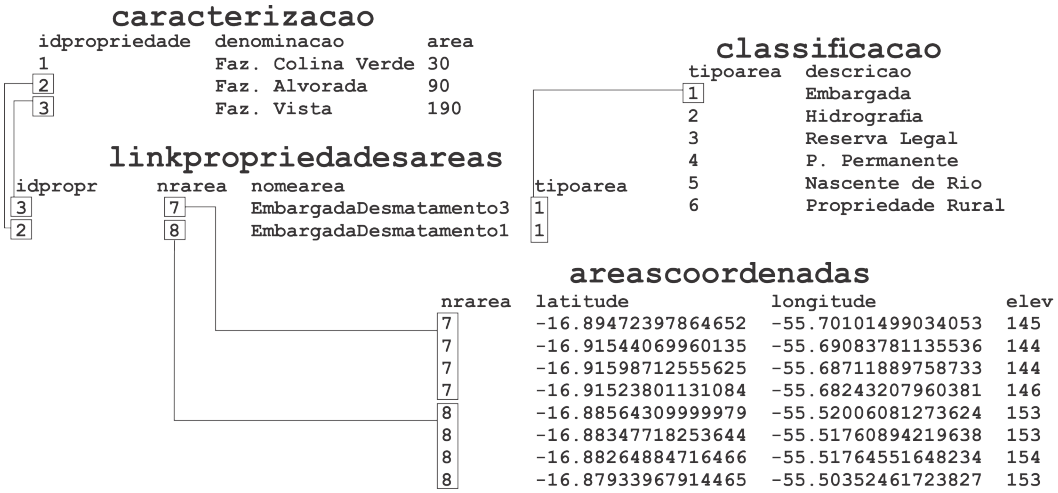

Figura 8. Associação de Área.

A construção do *dataset* também consiste em atribuir conteúdos de forma supervisionada para agregar valores aos dados finais. A associação das coordenadas extraídas permite que sejam realizadas análises como distanciamento de pontos de interesse e identificação de áreas relevantes como nascente de rios e áreas urbanas. Os dados resultantes estão disponíveis no repositório `https://github.com/clovissjunior/Geo`. Quanto ao arquivo com as coordenadas disponibilizado no artigo, trata-se de um exemplo, portanto o tamanho é reduzido com poucos KBs. Nessa fase da pesquisa optou-se por um formato genérico como o *csv* em razão da portabilidade entre ferramentas, pois o formato permite a importação dos dados por grande parte das ferramentas de geoprocessamento. Outros formatos de arquivos de dados georreferenciados como o padrão aberto geoPackage[2] poderá ser utilizado em nova versão da pesquisa. Outro ponto importante refere-se ao formato no qual as coordenadas estão armazenadas, o padrão utilizado é o *default* do *Google* Earth, apesar da elevação ter sido incluída no arquivo de dados nessa fase da pesquisa não foi utilizada. As elevações foram obtidas com a *API openElevation*[3] com propósito de agregar recursos para aplicações com demandas de plotagem de relevos, como identificação de erosões, alargamento e estreitamento de inconstas em leitos de rios. A Tabela 1 mostra dados reais do dataset gerado a partir da extração de dados de arquivos KML.

[2]http://www.geopackage.org/

[3]https://open-elevation.com/

Área	Latitude	Longitude	Elevação
AreaPreservacao	-16.88181703145399	-55.50015727578038	154
AreaPreservacao	-16.88564309999979	-55.52006081273624	153
UBA	-16.91598712555625	-55.68711889758733	144
AreaPlantio	-16.9073124506972	-55.67448939836068	150
AreaPreservacao	-16.87329153499706	-55.52168141416891	154
UBS	-16.9073124506972	-55.67448939836068	150
AreaPlantio	-16.91176189376817	-55.67624808991618	147

Tabela 1. *Dataset* Proposto com as Coordenadas Extraídas.

Outro ponto importante ainda associado à agregação de valores aos dados extraídos é abordado por [SCITEPRESS 2014] e consiste na categorização dos dados. Isso possibilita análises comparativas em diferentes regiões rurais como áreas protegidas, parques nacionais e áreas de destaque natural entre outras. Nesse contexto, é relevante incluir dados complementares, isso agrega recursos que posteriormente são utilizados para melhorar a geração de informações.

4.4. Utilização dos Dados Gerados

O distanciamento entre áreas rurais ou pontos de interesse próximos a áreas urbanas são exemplos de possibilidades reais para uso do *dataset* resultante. Para ilustrar essa aplicação foi utilizada uma base de dados complementar disponibilizada pelo governo federal com a localização geográfica central de todos os municípios e distritos do Brasil. Essa referência foi utilizada para calcular a distância entre áreas de interesse contidas no *dataset* proposto e coordenadas disponibilizadas no *dataset* governamental. A estrutura complementar é composta pelos seguintes atributos: distrito, município, micro região, meso região, estado, categoria, localidade, longitude, latitude, altitude e tipo de localização (rural ou urbano) e está disponível no repositório `https://github.com/clovissjunior/Geo`. A Tabela 9 mostra um fragmento do dataset complementar utilizado para validar o dataset proposto.

Distrito	Município	Microrregião	Mesoregião	UF	Categoria	Localidade	Longitude	Latitude	Altitude	Tipo
Três Rios	Três Rios	Três Rios	Cen Fluminense	RJ	Cidade	Três Rios	-43.2116094443354	-22.1174480009517	274.132	Urbano
Bemposta	Três Rios	Três Rios	Cen Fluminense	RJ	Vila	Bemposta	-43.0983728519305	-22.1401589383516	331.745	Urbano
Valença	Valença	Barra Do Piraí	Sul Fluminense	RJ	Cidade	Valença	-43.7048436783271	-22.2447194311794	550.859	Urbano
Parapeúna	Valença	Barra Do Piraí	Sul Fluminense	RJ	Vila	Parapeúna	-43.8299830430942	-22.0935196372014	445.620	Urbano
Pentagna	Valença	Barra Do Piraí	Sul Fluminense	RJ	Vila	Pentagna	-43.7529098449965	-22.1607158645465	474.021	Urbano

Figura 9. Dataset Governamental Complementar.

Uma aplicação ilustrativa utilizando a equação de ponto médio foi utilizada para validar as coordenadas extraídas no *dataset* proposto, mostrando distanciamentos entre pontos *shape* de uma área geográfica. Conforme mostra a equação 1, a fórmula evita erros de arredondamento quando comparado a outros algoritmos, como distância de Pitágoras Equirretangular ou Euclidiana que também é usado para calcular a distância entre duas coordenadas [Theoson et al. 2020].

$$hav(\Theta) = hav(\varphi 2 - \varphi 1) + \cos(\varphi 2) hav(\lambda 2 - \lambda 1) \quad (1)$$

Município	Localização	Distância	Mesoregião	Tipo	Estado
Barão de Melgaço	Barão de Melgaço	84 km	Centro-Sul	Urbano	MT
Barão de Melgaço	Joselândia	67 km	Centro-Sul	Urbano	MT
Santo Antonio	Mimoso	80 km	Centro-Sul	Urbano	MT
Santo Antonio	Vila Indigena Terena	61 km	Centro-Sul	Urbano	MT

Tabela 2. Resultados para Consulta Ilustrativa.

Complementarmente, foi utilizada uma equação para calcular ponto central para áreas geográficas a partir de polígonos apresentado na equação 2.

$$\Delta\sigma = (\sin_1 \sin_2 + \cos_1 \cos_2 \cos(\Delta\lambda)) \tag{2}$$

Conforme apresentado na equação, as informações geradas com cálculos de distâncias entre áreas permite verificações de proximidades para fins diversos como áreas de preservação, nascentes de riso etc.

A Tabela 2 apresenta de forma tabular as informações de algumas localidades encontradas em um raio específico.

Foram encontradas algumas dificuldades quanto à criação de um parser para realizar as análise dos conteúdos capturados em razão da heterogeneidade dos mesmos. Apesar das páginas seguirem uma mesma estrutura, algumas podem apresentar uma estrutura específica dependendo da ferramenta utilizada para criá-las. Os conteúdos relacionados com dados georreferenciados apresentados no artigo podem ser utilizados realização de análises de distanciamento de áreas degradadas no domínio ambiental e delimitação de áreas plantadas ou para planejamento de plantios no domínio agrícola. A disponibilização desses dados em ambiente Web é grande e representa uma importante fonte dados. Outro ponto importante refere-se à disponibilização de dados georreferenciados por órgãos governamentais que também contribui como complementos alternativos para a criação de *datasets*. É importante destacar que o artigo está em uma fase inicial e validações ainda serão realizadas para identificar possíveis enriquecimento a serem realizados de forma a complementar o *dataset* final.

5. Conclusões e Trabalhos futuros

O artigo mostrou uma alternativa para a criação de *datasets* com dados georreferenciados aplicado à agricultura e meio ambiente. A proposta principal é apresentar uma alternativa para uma demanda constante por dados nos domínios supracitados. O custo para obtenção deste tipo de dado é relativamente alto, o uso de ferramentas sem licenciamento comercial mostra-se uma alternativa viável conforme demonstrado no artigo. Como trabalhos futuros propõe-se estender os recursos de importação a outros tipos de fontes de dados web além da estrutura de arquivos apresentada neste artigo. Isso contribuirá com mais recursos agregados às propriedades rurais no contexto ambiental. Outro ponto importante é a extensão dos critérios utilizados para importação outros tipos de arquivos de dados georreferenciados como *SHP*, ampliando as possibilidades de criar ou agregar dados a *datasets* nesse domínio.

Referências

Azad, S., Wasimi, S., and Ali, A. (2018). Business data enrichment: Issues and challenges. In *Business Data Enrichment: Issues and Challenges*, pages 98–102.

Azeroual, O. and Jha, M. (2021). Without data quality, there is no data migration. *MDPI*, 5(2):24.

Bonamigo, A. (2015). Impactos na adequação das áreas de preservação permanente de imóveis rurais ao disposto na lei nº 12.651 e lei nº 4.771 (código florestal).

Cruz, I. F., Ganesh, V. R., and Mirrezaei, S. I. (2013). Semantic extraction of geographic data from web tables for big data integration. In *Proceedings of the 7th Workshop on Geographic Information Retrieval*, GIR '13, page 19–26, New York, NY, USA. Association for Computing Machinery.

Dong, X. L., Hajishirzi, H., Lockard, C., and Shiralkar, P. (2020). Multi-modal information extraction from text, semi-structured, and tabular data on the web. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '20, page 3543–3544, New York, NY, USA. Association for Computing Machinery.

Gong, D., Wang, D. Z., and Peng, Y. (2017). Multimodal learning for web information extraction. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 288–296, New York, NY, USA. Association for Computing Machinery.

Imbrenda, V., Calamita, G., Coluzzi, R., D'Emilio, M., Lanfredi, M., Perrone, A., Ragosta, M., and Simoniello, T. (2013). Free and open source software for land degradation vulnerability assessment. *None*, page 11153.

Jaya, I., Sidi, F., Ishak, I., Affendey, L., and A. Jabar, M. (2017). A review of data quality research in achieving high data quality within organization. *Journal of Theoretical and Applied Information Technology*, 95:2647–2657.

Lloret-Gazo, J. (2020). A browserless architecture for extracting web prices. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, SAC '20, page 2193–2200, New York, NY, USA. Association for Computing Machinery.

openforis (2021). Open foris. http://openforis.org/. (Accessed on 10/19/2021).

SCITEPRESS (2014). Database design of a geo-environmental information system. In *Proceedings of the 16th International Conference on Enterprise Information Systems*. SCITEPRESS - Science and and Technology Publications.

Theoson, L., Anthony, R., and Purnama, J. (2020). Distance-measurement-decision-making backend system using nodejs. In *Proceedings of the International Conference on Engineering and Information Technology for Sustainable Industry*, ICONETSI, New York, NY, USA. Association for Computing Machinery.