COLETA DE DADOS DE IMÓVEIS DE FORMA AUTOMATIZADA PARA FINS DE POLÍTICAS PÚBLICAS

Property data collection automated for public policy purposes

Caroline Bernardo Silva
Universidade Federal de Santa Catarina
carolinebernardosilva@gmail.com

Eduardo Schmidt Longo
Universidade do Estado de Santa Catarina
eduardosl.geo@gmail.com

Everton da Silva
Universidade Federal de Santa Catarina
everton.silva@ufsc.br

Resumo:
O presente artigo aborda a possibilidade da coleta de dados de imóveis pelo Poder Público, através do método de automatização chamado de web crawler. A pesquisa partirá da premissa de que, entre diversas variáveis e dados que podem ser utilizados para as políticas públicas, um deles seriam aqueles que descrevem de alguma forma a espacialização de imóveis em um recorte geográfico. Todavia, a forma tradicional de coletar estes dados seria a mecânica, individualizando cada extração de informação que interessasse. Desta forma, este estudo preceitua o método de raspagem automatizada como uma forma de otimizar tarefas aos planejadores e tomadores de decisão.

Palavras-chave: web crawler; LGPD; coleta de dados.

Abstract
This article approach the possibility of collecting real estate data by the Government, through the automation method called a web crawler. The research is based on the premise, among several variables and data that can be used for public policies, one of them would be those that somehow describe the spatialization of properties in a geographic cut. However, the traditional way of collecting this data would be mechanics, individualizing each extraction of information that interested. Thus, this study provides the automated scraping method as a way to optimize tasks for planners and decision makers.

Keywords: web crawler; LGPD; data collect.

1. INTRODUÇÃO

Na busca por equidade em políticas públicas é importante o conhecimento do local de interesse em que serão aplicadas, um cadastro territorial provido com informações sobre a distribuição e disponibilidade do solo é essencial para definição dessas políticas públicas (ERBA, PIUMETTO, 2016). Para tal, é fundamental que sejam obtidos dados que auxiliem em análises estratégicas e embasem suas aplicabilidades pelo Poder Público. Todavia, apesar da abundância de dados disponíveis, é comum se deparar com dificuldades de obtenção e mesmo manutenção desses dados, recorrendo a pesquisas manuais, pontuais e sem retenção dos dados para uso em outros estudos.

Uma linha de política pública, voltada ao planejamento territorial, é a de solo, que necessita de dados com informações sobre o mercado imobiliário - preço dos imóveis, área e outras características que impactem na formação do valor dos imóveis. Neste sentido, esse artigo enfoca-se na importância e legalidade da coleta automatizada de dados imobiliários para realização de políticas públicas voltadas ao solo, como: mais-valias, IPTU progressivo, outorga onerosa e estudos do mercado imobiliário que são geram captação de recursos e auxiliam na capacidade dos municípios de realizar suas atribuições (DE CESARE et al., 2015).

Será abordado na análise o método de coleta utilizando web crawlers e levantadas reflexões sobre sua legitimidade diante a Lei Geral de Proteção de Dados Pessoais (LGPD) no uso pelo Poder Público. De forma que os dados coletados com o uso da ferramenta, sejam armazenados, utilizados, e que seu uso/exposição respeite a anonimização das pessoas relacionadas aos imóveis e eventos de marcado, bem como promova benefícios à gestão pública.

No cenário atual, onde a dinâmica de transformação do solo e da sociedade é cada vez mais intensa e pautada em informações, é vital para administração pública contar com dados que possam embasar suas análises e ações visando o bem estar das pessoas e uma cidade/município sustentável. Para tanto, o Poder Público deve apoiar-se nos instrumentos de política e nas boas práticas de outros entes, que invariavelmente se materializam a partir de informações urbana, a fins de respeitar e aplicar as diretrizes preconizadas pelo Estatuto da Cidade (SAULE JR, N., ROLNIK, R., 2001). Daí a importância de se discutir o tema proposto por este artigo, de modo a possibilitar o aparelhamento metodológico das administrações públicas.

2. WEB CRAWLERS COMO MÉTODO DE COLETA

Para obter dados de imóveis é comum o pesquisador se utilizar de sites de imobiliárias e agregadores de anúncios, realizando uma coleta manual onde são obtidas as informações de forma direcionada e adicionadas ao banco de dados por digitação. Esse tipo de pesquisa é baseado na coleta das informações que estão disponíveis nos sites, ou seja, dados que já estão abertos pelo administrador. Neste procedimento raramente sobrecarrega-se o serviço de administração web, uma vez que é realizado de forma lenta e espaçada.

Em tese, uma vantagem dos dados coletados manualmente é sua verificação/validação no momento da coleta, facilitando sua adição ao banco de dados devido ao saneamento prévio, o que possibilita o imediato manuseio dos dados e o desenvolvimento de processamentos e análises.

Todavia, é frequente o banco de dados não receber manutenção e a coleta iniciar do zero sempre que haja demanda por uma pesquisa. Essa prática prejudica estudos que envolvam dados temporais dos imóveis, como: variação dos preços dos imóveis ao longo do tempo, análise do valor do solo após implementação de obras de infraestrutura para futura cobrança de mais-valias e de contribuição de melhoria, por exemplo. A sistematização dessa coleta e a alimentação do banco de dados de forma periódica gera amparos para implementação de políticas públicas e estudos sobre o espaço territorial.

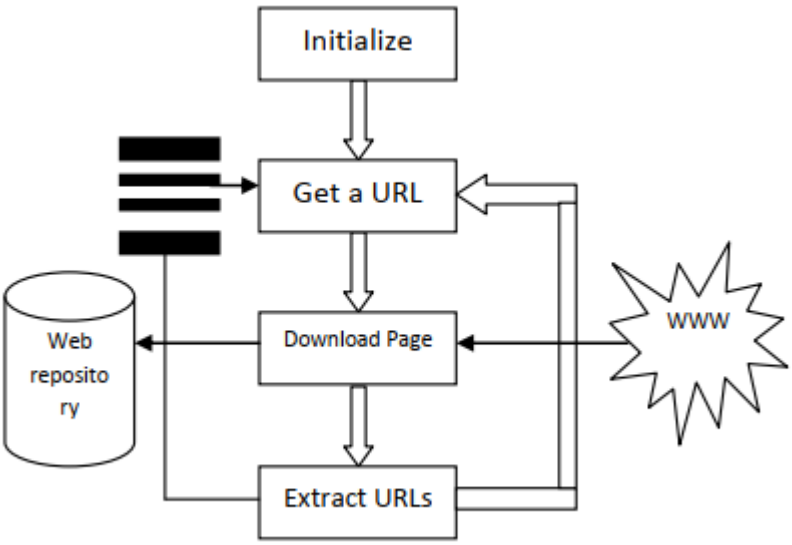
Uma solução para manutenção de dados de imóveis de forma constante e abundante é a coleta de automatizada de dados através da utilização de web crawlers, que são programas/softwares ou scripts que realizam leituras de websites, consumindo seu conteúdo de

forma metódica e automatizada (KAUSAR; MCGUFFEE, 2013) e realizando uma coleta de massiva de dados.

No geral, os web crawlers envolvem duas técnicas, uma com propósito geral de armazenar websites para indexação, geralmente são *crawlers* mais lentos devido a grande quantidade de processamento, e outra com propósito focado, onde são coletados dados predefinidos (KAUSAR; MCGUFFEE, 2013), sendo esse utilizado com intuito requisição e armazenamento de dados. Esta última foi a que serviu como eixo para o desenvolvimento do artigo.

Na prática o *web crawler* focado realiza consulta no servidor *web*, requisita os dados, em HTML ou outras formas de composição web, e efetua o *parse* (captura e transformação) para extrair as informações (MITCHELL, 2019). Essa extração gera dados que podem ser armazenados em diversos tipos de arquivos de saída ou exportados diretamente para o banco dados.

Figura 1- Esquema de funcionamento do web crawler



Fonte: KAUSAR; MCGUFFEE, 2013

3. PERSPECTIVAS SOBRE O USO DE WEB CRAWLER E A PROTEÇÃO DE DADOS NO BRASIL

A Lei Nº 13.709, DE 14 DE AGOSTO DE 2018, conhecida como Lei Geral de Proteção de Dados Pessoais (LGPD), define-se como responsável por trazer ao universo jurídico brasileiro conceitos e premissas sobre tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, conforme preconiza seu artigo 1º. Possui como objetivo a proteção dos direitos fundamentais da liberdade e da privacidade, além do livre desenvolvimento da personalidade da pessoa natural quanto ao que tange a proteção de seus dados.

Desta forma, compreende-se que a LGPD empenha-se especificamente quanto

ao tratamento de dados pessoais (SANTOS, 2019), não sobrepondo-se a priori em dados de pessoa jurídica, ou documentos sigilosos/confidenciais, dados de negócios (mercado), algoritmos, fórmulas, softwares, patentes, entre outros documentos ou informações que não sejam relacionadas a pessoa natural identificada ou identificável (MALDONADO, 2019).

De antemão, implica-se na lembrança de que estas informações ou documentos supracitados possuem sua tutela em diversos outros diplomas legais, como na Lei de Dados Abertos, Lei de Propriedade Industrial (Lei 9.279/1996), a Lei de Direitos Autorais (Lei 9.610/1998), a Lei de Software (Lei 9.609/1998), entre diversos outros, cabendo a cada situação e contexto sua sensível análise e remediação. Todavia, se algum dos casos não abarcados explicitamente pela LGPD contenha dados pessoais identificáveis (personalidade e/ou sujeito) estarão logo protegidos por aquela concomitantemente.

Fator preponderante de preocupação e reflexão crescente nos últimos tempos é a atual capacidade de processamento das máquinas, intensificando a coleta, armazenamento, tratamento e compartilhamento de dados cada vez mais desejados e importantes para incontáveis finalidades. Portanto, a LGPD é um processo de precaução do país quanto a perspectiva jurídica dos dados.

Dentre várias definições que a LGPD traz no bojo de seu artigo 5º, esta pesquisa prende atenção àquelas relacionadas aos conceitos de dado pessoal, tratamento, e anonimização:

> I - dado pessoal: informação relacionada a pessoa natural identificada ou identificável;
>
> X - tratamento: toda operação realizada com dados pessoais, como as que se referem a coleta, produção, recepção, classificação, utilização, acesso, reprodução, transmissão, distribuição, processamento, arquivamento, armazenamento, eliminação, avaliação ou controle da informação, modificação, comunicação, transferência, difusão ou extração;
>
> XI - anonimização: utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo (BRASIL, 2018);

Tal abordagem com enfoque nestas três expressões se dará a partir do escopo da reflexão levantada, a *coleta de dados imobiliários de forma automatizada para fins de políticas públicas*.

Discorrer sobre coleta de dados e Lei Geral de Proteção de Dados pode se tornar algo amplo e diversificadamente caótico caso não se defina alguns atributos sobre tal atividade, como *tipo de dado, a finalidade da coleta e quem são os agentes envolvidos* - e definir isto pode incorrer como boas práticas da coleta de dados pelo Poder Público, harmonizando e assegurando a atividade de busca pelo dado (com função à sociedade) em relação aos seus controladores e proprietários.

Desta feita, salienta-se que o escopo e objetivo geral da coleta de dados analisada neste momento se dará onde o Poder Público, na atribuição de suas atividades, necessidades e finalidades, carece de dados diversificados que podem transformarem-se em informações de retorno positivo às políticas de solo. De outro lado, reforça-se que neste mesmo momento, há

outros agentes de transformação do solo (CORREA, 1993) que podem conter estes dados de interesse ao município, como por exemplo, os agentes imobiliários.

Logo, uma vez que o interesse do Poder Público é coletar apenas os atributos das unidades imobiliárias, no que tange à informações despersonalizadas, ou seja, que não possibilitem qualquer associação entre sujeito e dado, mostra-se útil e imperioso a existência de legitimidade daquela para com o uso de web crawlers - conforme preconiza o próprio art. 5º, incisos I, X, e XI da LGPD.

Outro ponto que merece destaque diz respeito ao Art. 7º, incisos III e IV, com a premissa do respeito ao Princípio da Anonimidade, combinado com parágrafos 3º e 4º:

> Art. 7º O tratamento de dados pessoais **somente poderá ser realizado** nas seguintes hipóteses:
>
> III - pela **administração pública**, para o tratamento e uso compartilhado de dados **necessários à execução de políticas públicas previstas em leis** e regulamentos ou respaldadas em contratos, convênios ou instrumentos congêneres, observadas as disposições do Capítulo IV desta Lei;
> IV - para a realização de estudos por órgão de pesquisa, garantida, sempre que possível, **a anonimização dos dados pessoais**;
> [...] § 3º O tratamento de dados pessoais **cujo acesso é público** deve considerar a finalidade, a boa-fé e o interesse público que **justificaram sua disponibilização**.
> § 4º É dispensada a exigência do consentimento previsto no caput deste artigo **para os dados tornados manifestamente públicos pelo titular, resguardados os direitos do titular** e os princípios previstos nesta Lei.
> (BRASIL, 2018. Grifo nosso).

Analisando de forma fragmentada cada trecho, pode-se arguir que a administração pública, baseada nos pressupostos das políticas de solo, poderiam ser beneficiados pelo uso dos dados em detrimento dos agentes imobiliários, fazendo jus do art. 7º inciso III como justificativa para requerimento do tratamento.

Na mesma esfera, através de determinado órgão específico para garantia das pesquisas para políticas públicas, e consequentemente a necessidade de coleta de dados aproveitáveis, o respeito ao princípio da anonimização dos dados pessoais pela entidade de pesquisa coletadora legitima o uso daqueles.

Ademais, pressupõe-se a partir da própria LGPD que, sendo os dados manifestadamente públicos por vontade e interesse do titular, e sempre resguardados os direitos fundamentais daquele (MENDES, 2014), estando seu acesso público, o Município no uso de suas prerrogativas com finalidade de políticas de interesse público e munido de boa-fé encontraria respaldo também na LGPD para busca e uso de determinados dados imobiliários.

4. BOAS PRÁTICAS DE USO DE WEB CRAWLER PELO PODER PÚBLICO

Muitos sites tendem a bloquear *bots* para evitar a busca e coleta de suas informações, mas os bots são adaptados conforme surgem as barreiras e utilizam técnicas/scripts para se camuflarem e realizar a obtenção de dados sem serem barrados por esses recursos. Sendo que um dos maiores desafios dos sites para realizar esse bloqueio é a diferenciação entre os bots e humanos realizando coletas e acessos manuais, por isso, uma das técnicas utilizadas é camuflar o bot para simular comportamento humano.

Entre as técnicas estão a utilização de temporizadores para que as requisições sejam feitas de forma mais espaçada, uso de cabeçalhos com identificação falsa e/ou randomizados a cada raspagem, não requisitar dados em campos ocultos e mudança constante de IP (Internet Protocol).

Apesar das técnicas serem eficientes do ponto de vista da coleta de dados, elas ferem as boas práticas de uso e reforçam o estigma de que web crawlers geram ônus aos sites em que são realizadas as raspagens. Para evitar que o crawler seja banido de uma página e trabalhe, é aconselhado que seja implementado de forma transparente.

Entre as boas práticas, pode-se citar:

a - utilização de um formulário claro, com as informações da origem do web crawler, como o setor do órgão público e e-mail para contato;

b - uso de temporização nas requisições de forma a não sobrecarregar o servidor de administração web;

c - realizar as requisições de preferência em horários não comerciais, ou seja, que possuem baixa demanda pelo usuário; e

d - o conhecimento dos “termos de serviço” do controlador dos dados, buscando perceber se eles proíbem qualquer tipo de extração de dados.

Figura 2 - Exemplo de requisição do web crawler

scrapy > ouv > settings.py > ...

```
1   # -*- coding: utf-8 -*-
2
3   # Scrapy settings for ouv project
4
5   BOT_NAME = 'Identificação do coletor'
6
7   SPIDER_MODULES = ['ouv.spiders']
8   NEWSPIDER_MODULE = 'ouv.spiders'
9
10  # Crawl responsibly by identifying yourself (and your website) on the user-agent
11  USER_AGENT = 'Idenficação do usuário'
12
13  # Configure maximum concurrent requests performed by Scrapy (default: 16)
14  CONCURRENT_REQUESTS = 4
15
16  # Configure a delay for requests for the same website (default: 0)
17  # See https://doc.scrapy.org/en/latest/topics/settings.html#download-delay
18  # See also autothrottle settings and docs
19  DOWNLOAD_DELAY = 2
20
```

Fonte: Elaborado pelo autor (2020).

Além disso, deve-se sempre considerar entrar em contato com o administrador do site para informar as suas finalidades e que tipo de dados estão sendo coletados, especialmente na

demonstração de não haver possibilidade de desrespeito ao princípio da anonimidade ou qualquer ligação entre pessoa titular do dado e o dado em si, e porventura negociar uma contrapartida em seu uso.

5. CONCLUSÕES

O presente artigo teve como objeto de análise o método automatizado de coleta de dados baseado em web crawlers, considerando-o como ferramenta funcional e eficiente na busca por dados imobiliários.

No escopo e descrição do método foi evidenciada a necessidade que o Poder Público possui em obter acesso à dados imobiliários, e consequentemente informações, especialmente no que tange às suas atribuições de Estado em prover as melhores condições possíveis de vida aos cidadãos, tomando por ferramentas os instrumentos de política de solo e outras estratégias governamentais.

Apontou-se conceitualmente o que é o web crawler, suas possibilidades e características como mecanismo de coleta de dados, abordando-se, através das referências, seu funcionamento e operacionalização como ferramenta de apoio às ações da administração pública.

Após tais revisões, a pesquisa apresentou as considerações a respeito das perspectivas sobre o uso de Web Crawler quanto a proteção de dados no Brasil, especialmente no que diz respeito às expectativas sobre a Lei Geral de Proteção de Dados (LGPD).

A partir das compreensões sobre a LGPD em relação ao método automatizado de coleta de dados abordado, o que se vislumbra é a possibilidade do uso do web Crawler para fins de políticas públicas, respeitados concomitantemente o princípio da anonimização do proprietário dos dados, onde não é possível a vinculação daquele(s) ao objeto abordado/coletado, além de outras nuances que também ratificam a possibilidade do ato estudado.

Ademais, em decorrência de observações trazidas pela bibliografia do web crawler, e pensadas juntamente do raciocínio jurídico e ético-social do tema abordado, a pesquisa aduz sobre a recomendação pelo uso do que chamaremos de *boas práticas de coleta de dados por web crawler*, onde o responsável por fazer a coleta realiza uma abordagem transparente e uso de formas não-prejudiciais a qualquer um dos atores da relação com os dados.

Referências

BRASIL. Lei n. 13.079, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, 2018.

CORREA, Roberto Lobato. O espaço urbano: notas teórico-metodológicas. Geosul, v. 8, n. 15, p. 13-18, 1993.

DE CESARE, Cláudia M.; FERNANDES, Cintia Estefânia e Cavalcanti, CAROLINA Baima (Org). Imposto sobre a Propriedade Predial e Territorial Urbana: Caderno Técnico de Regulamentação e Implementação / Ministério das Cidades, 2015, p. 73.

ERBA, D. A.; PIUMETTO, M. A. Para leer el suelo urbano, Catastros multifinalitarios para la planificación y el desarrollo de las ciudades de América Latina, ano 2016, p.23.

KAUSAR, Abu; DHAKA, V.S.; SING, Sanjeev Kumar. Web Crawler: A Review. International Journal of Computer Applications, Jaipur, India, ano 2013, v. 63, n. 2, p. 31-36, 31 jan. 2013.

MALDONADO, Viviane Nóbrega; BLUM, Renato Opice; BORELLI, Alessandra. LGPD: Lei geral de proteção de dados: comentada. Revista dos Tribunais, 2019.

MENDES, Laura Schertel. Privacidade, proteção de dados e defesa do consumidor: Linhas gerais de um novo direito fundamental, p. 176. São Paulo: Saraiva, 2014.

MITCHELL, Ryan. Web Scraping com Python: Coletando mais dado na web moderna. 2. ed. São Paulo,SP: Novatec, 2019.

MYERS, Daniel; MCGUFFEE, James W. Chossing Scrapy. Jornal of Consortium for Computing Sciences in Colleges, [s. l.], ano 2015, p. 83-89, 31/10/2015.
ROBERT L. K. Tiong. Risks and Guarantees in BOT Tender. Journal of Construction Engineering and Management. ASCE. Vol. 121, Singapore,1995.

SANTOS, Dhiulia de Oliveira. A validade do consentimento do usuário à luz da lei geral de proteção de dados pessoais (Lei n. 13.709/2018). 2019.

SAULE JR, N.; ROLNIK, R. Estatuto da Cidade: guia para implementação pelos municípios e cidadãos. Pólis Instituto de Estudos Formação e Assessoria em Políticas Sociais e Caixa Econômica Federal, apoio Comissão de Desenvolvimento Urbanos da Câmara dos Deputados, Secretaria Especial de Desenvolvimento Urbano da Presidência da República, Câmara dos Deputados Brasília, 2001, p. 33.