

Igor Rocha: Protocolos de exclusão de robôs e a Justiça digital

01/01/2023

Há alguns meses escrevi um texto na revista eletrônica **Consultor Jurídico** apresentando a figura dos *web crawlers* e dos *web scrapers* [1] para os operadores do Direito, tendo, naquela ocasião, ressaltado que, apesar de bastante populares, tais ferramentas ainda estavam envoltas em polêmicas jurídicas que precisavam ser melhor estudadas.



Com o objetivo de dar continuidade a esta linha de

raciocínio, escrevo o presente texto para abordar uma dessas polêmicas, qual seja: a existência dos chamados protocolos de exclusão de robôs, também chamados de protocolos "*robot.txt*", dentro dos sítios digitais de órgãos estatais, em especial, dos tribunais de justiça no Brasil.

Para entender essa questão, porém, primeiro é necessário compreender o que são os ditos protocolos "*robot.txt*".

Nesse sentido, Maurice Schellekens [2] define tais protocolos como arquivos simples em formato de texto que podem ser inseridos dentro do código fonte de um determinado site com o intuito de informar a futuros programadores se tal sítio digital aceita, ou não, que *web crawlers* e *web scrapers* acessem os seus conteúdos.

Ou seja, tais arquivos são uma forma que o desenvolvedor de determinado site tem para dizer que robôs não são bem vindos no seu site bem como para estipular os limites de acesso que eles terão.

Esses protocolos se mostram como uma ferramenta bastante poderosa pois permitem que os desenvolvedores controlem o quanto as informações de seus sites poderão ser analisadas, parametrizadas ou mesmo publicizadas por meio de robôs de terceiros.

Assim, por exemplo, caso um desenvolvedor não queira que o conteúdo de seu site seja amplamente divulgado, poderá impedir a indexação de algumas de suas páginas no mecanismo de busca do Google via a inclusão de um arquivo que impeça que os *crawlers* ou *scrapers* da

empresa do Vale do Silício o acessem.

No mesmo sentido, caso não queira que as informações de seu site sejam analisadas e parametrizadas para fins estatísticos, poderá igualmente inserir a referida espécie de arquivo para dificultar a coleta automatizada de dados no seu site e assim por diante.

Ressalta-se aqui, entretanto, que a inclusão de tal espécie de arquivo nos sites não é obrigatória de modo que a decisão de incluir eles se trata de uma escolha discricionária do desenvolvedor, motivada pelas razões que considerar mais corretas.

A questão, porém, passa a interessar muito o direito quando se analisa as diversas consequências e problemas que surgem a partir do momento que a utilização de arquivos "*robot.txt*" começa a ocorrer também no âmbito dos sítios digitais ligados a órgãos públicos.

Isso ocorre pois, neste cenário, tal discussão toma contornos mais sérios, ligados ao dever dos órgãos públicos de publicizarem suas informações, a forma como o irão fazer bem como acerca da necessidade de se garantir o acesso amplo as bases de dados públicas.

A questão é complexa e envolve nuances que extrapolam os limites do presente texto, mas é interessante apontar que alguns pesquisadores, estudando a forma como deve ser montada uma política de acesso aos dados públicos brasileiros, já abordaram a temática e mesmo encontraram sítios digitais públicos no Brasil que se utilizam de tais protocolos.

Nesse sentido, Marcel Hofling [3], em sua dissertação de mestrado, apontou que, no âmbito do Tribunal de Justiça de São Paulo, mais especificamente dentro do seu sistema de processo eletrônico, é possível verificar a presença de protocolo desautorizando o acesso de qualquer robô a pasta "cjsg", local onde se encontram as decisões judiciais.

Tal situação se mostra bastante prejudicial para toda a sociedade pois, ao impedir que mecanismos automatizados acessem tais dados sem que exista qualquer outra rota de acesso para a grande quantidade de informações existentes nos bancos de dados públicos, cria-se um óbice desnecessário de acesso à dados que, em sua origem, são públicos.

Inspirado por essa experiência venho conduzindo um estudo que tem por intuito coletar e analisar informações acerca da presença de arquivos "*robot.txt*" nos sites dos Tribunais de Justiça ao redor do país.

Embora ainda não tenha sido concluído, tal pesquisa já apresentou alguns achados interessantes que vão na mesma linha das descobertas realizadas pelos pesquisadores citados, mas que, por expandirem a amplitude do problema encontrado, merecem serem compartilhados [4].

O primeiro desses achados remete ao Tribunal de Justiça de Tocantins onde a análise das páginas do sistema de processo judicial eletrônico de primeira e de segunda instância (Eproc) permitiram descobrir a existência de protocolos que desautorizam o acesso de qualquer robô a qualquer uma das páginas do referido sistema, sem qualquer exceção

Em sentido parecido, a análise da página inicial do site do Tribunal de Justiça de Minas Gerais permitiu encontrar um protocolo que desautoriza o acesso de qualquer robô, exclusivamente, a sessão do site que contém informações sobre os ex-presidentes do tribunal.

Por fim, temos o caso do site do Tribunal de Justiça do Distrito Federal e Territórios onde, a partir do estudo realizado, foi possível encontrar um protocolo que, até o momento, talvez seja o mais interessante de todos pois desautoriza o acesso de qualquer robô a algumas páginas do site do referido tribunal, como as páginas que contém informações acerca dos concursos anteriores para o provimento dos cargos de magistrado, analista e técnico do tribunal.

A parte mais interessante, entretanto, está no fato de que, além disso, o referido protocolo desautoriza, especificamente, o acesso dos robôs do *Google* a uma outra grande parcela do site.

Da análise da coletânea de casos acima elencados é possível perceber que a presença dos protocolos de exclusão de robôs já é uma realidade dentro dos sites de muitos Tribunais de Justiça no Brasil o que, inevitavelmente, faz com que passem a surgir uma gama de questionamentos, como:

- Quantos são os tribunais que contam com protocolos de exclusão de robôs dentro seus códigos-fonte?
- Qual a amplitude de restrição de acesso aos dados públicos que a existência de tais protocolos oferece?
- Por quais motivos alguns tribunais optam por inserir os arquivos “robot.txt” em seu códigos-fonte?
- Quais são as consequências da utilização de tais protocolos em sites públicos no Brasil?
- Qual o impacto da utilização de tais protocolos para a formação de uma política de dados abertos brasileira?

Inegavelmente o tema se mostra controverso e cheio de nuances que ainda precisam ser exploradas pelos operadores do Direito brasileiro. A questão principal é que, com o aumento cada vez maior da importância da análise de dados para a prática do Direito, responder a essas e muitas outras dúvidas deixa de ser uma questão opcional e passa a se tratar de uma questão obrigatória.

Referências

HÖFLING, Marcel de Souza. INTELIGÊNCIA ARTIFICIAL NO ESTUDO DO DIREITO PROCESSUAL: metodologia de pesquisa empírica na agenda do Acesso à Justiça. 2022. 1-177 f. Dissertação de Mestrado — Universidade de São Paulo, São Paulo, 2022.

ROCHA, Igor Moraes. Web crawlers, web scrapers e a sua importância para o Direito. Consultor Jurídico, São Paulo, 18 set. 2022. p. 1-1.

SCHELLEKENS, Maurice. *Robot.txt: balancing interests of content producers and content users*. Em: LEENES, Ronald; KOSTA, Eleni (org.). *Bridging distances in technology and regulation*. [S. l.]: Wolf Legal Publishers , 2013. p. 174-188.

[1] ROCHA, Igor Moraes. Web crawlers, web scrapers e a sua importância para o Direito. Consultor Jurídico, São Paulo, 18 set. 2022. p. 1-1.

[2] SCHELLEKENS, Maurice. *Robot.txt: balancing interests of content producers and content users*. Em: LEENES, Ronald; KOSTA, Eleni (org.). *Bridging distances in technology and regulation*. [S. l.]: Wolf Legal Publishers, 2013. p. 174-188.

[3] HÖFLING, Marcel de Souza. INTELIGÊNCIA ARTIFICIAL NO ESTUDO DO DIREITO PROCESSUAL: metodologia de pesquisa empírica na agenda do Acesso à Justiça. 2022. 1-177 f. Dissertação de Mestrado — Universidade de São Paulo, São Paulo, 2022.

[4] Importante ressaltar que tais resultados são preliminares, ou seja, podem passar por mudanças no momento de apresentação do resultado final a ser publicado em 2023, situação na qual a metodologia empregada no estudo também será minuciosamente explicitada.

Fonte: <https://conjur.jumps.com.br/2023-jan-01/igor-rocha-protocolos-exclusao-robos-justica/>