A raspagem de dados (web scraping) é uma maneira de extrair informações apresentadas nos sites. Com estes raspadores, também chamados de "bots", é possível coletar grandes quantidades de dados para reportagens. Por exemplo: o autor utilizou um raspador para fazer esta reportagem e comparar os preços do álcool entre Quebec e Ontário. Outro colega usou esta técnica para coletar de dados e comparar os preços de aluguel em vários bairros de Montreal com anúncios de Kijiji.

Mas quais são as regras éticas que os repórteres devem seguir durante a raspagem na web?

Essas regras são particularmente importantes, pois, para pessoas que não são nerds, a raspagem de dados na web parece uma invasão. Infelizmente, os códigos de ética não dão uma resposta clara a esta pergunta.

Dados públicos ou não?

Este é o primeiro consenso dos jornalistas de dados: se uma instituição publicar dados em seu site, esses dados são automaticamente públicos.

“Seja um humano que copia e cola os dados, ou um humano que cria um programa de computador para fazer isso, é o mesmo. É como contratar 1000 pessoas que trabalhariam para você. É o mesmo resultado”, afirma Cédric Sam, que trabalha para o South China Morning Post,

No entanto, os servidores do governo também hospedam informações pessoais sobre os cidadãos. Aqui está o limite muito importante entre raspagem na web e hackers: o respeito à lei.

Os repórteres não devem investigar dados protegidos. Se um usuário comum não puder acessá-lo, os jornalistas não devem tentar obtê-lo. Roberto Rocha, que até recentemente era repórter de dados do Montreal Gazette, acrescenta que os jornalistas devem sempre ler os termos e condições de uso do usuário para evitar problemas.

Outro detalhe importante a ser verificado: o arquivo robots.txt, que pode ser encontrado na raiz do site e que indica o que pode ser raspado ou não.

Identifique-se ou não?

O jornalista deve sempre se identificar como tal antes de fazer perguntas. Mas o que acontece no caso de um robô?

Glen McGregor, repórter de assuntos nacionais do Ottawa Citizen, acha que a mesma regra deve ser seguida. Ele diz que, para impedir que a pessoa que administra o site pense que está sendo invadido, no cabeçalho http, ele sempre coloca seu nome, seu número de telefone e uma mensagem dizendo quem ele é. Assim, caso surjam problemas, a equipe do site pode chamá-lo para solucionar dúvidas ou questões.

Nem todo mundo pensa o mesmo. Philippe Gohier, editor-chefe da L'Actualité, faz todo o possível para que não o identifiquem: às vezes, ele usa proxy, muda seu IP e seus cabeçalhos para parecer um humano navegando o site, ao invés de um um robô. Ele diz: “respeite as regras, mas permaneça anônimo”.

Para o autor do artigo, de certo modo, não se identificar ao extrair dados equivale a usar microfones ou câmeras escondidas ao fazer entrevistas.

O Código de Ética da FPJQ (Professional Federation of Quebec Journalists) possui as seguintes regras para justificar procedimentos de apuração secretos, que devem ser excepcionais:

Por exemplo, em casos onde “as informações solicitadas forem de interesse público definitivo ou em que ações socialmente repreensíveis devem ser expostas”. Ou ainda em casos onde a informação "não pode ser obtida ou verificada por outros meios, ou outros meios já foram utilizados sem sucesso"; ou ainda quando "o benefício público é maior que qualquer inconveniente para as pessoas".

Além disso, ele acrescenta que o público deve ser informado se algum desses métodos for usado para obter as informações.
Portanto, a melhor prática é sempre identificar o robô. Mas se houver o risco de a instituição ocultar informações, por exemplo, é melhor ser discreto sobre sua identidade. De qualquer forma, se houver receio de que o robô seja bloqueado, você poderá alterar facilmente seu endereço IP e esse problema será resolvido.
Outros jornalistas acham que é melhor pedir os dados primeiro e extraí-los se eles forem negados. O bom disso é que, se os dados são entregues, o repórter economiza muito tempo.

Publique seu código ou não?

A transparência é fundamental para os jornalistas terem credibilidade junto ao público.
A maioria dos repórteres é transparente sobre os dados em que suas histórias são baseadas.
No caso de códigos, se houver um erro na escrita, os dados obtidos podem levar a histórias completamente erradas.
Se for utilizado softwares de código aberto, é obrigatório revelá-lo para que outros o aprimorem e que sejam auditados.
Quando o código é escrito para fazer uma história jornalística, a decisão é mais complexa, porque esse código oferece vantagens sobre a concorrência. Portanto, o jornalista de dados Roberto Rocha acha que nem todos os códigos devem ser tornados públicos. Ele, como outros, tem um GitHub onde publica alguns.
Jean-Hugues Roy acha que os jornalistas devem compartilhar o código, assim como os cientistas compartilham suas metodologias, para ajudar a todos, mas há exceções: por exemplo, se você está trabalhando em um código que demorou muito tempo e não tem certeza se o publicará. Glen McGregor não o publica, mas o compartilha se alguém o solicitar.
Quando um repórter tem uma fonte, ele faz de tudo para protegê-la. Isso para ganhar sua confiança, mas também para mantê-la. O extrator de dados é como uma fonte jornalística.
Também é discutido se os códigos jornalísticos serão patenteados no futuro.
Por fim, um detalhe técnico relevante: “respeitar a infraestrutura da web é, obviamente, outra regra de ouro da extração da web. Sempre deixe alguns segundos entre suas solicitações e não sobrecarregue os servidores.”