Link: https://gijn.org/2015/08/12/on-the-ethics-of-web-scraping-and-data-journalism/

Llamamos “Web scraping” a un modo de extraer información de sitios web y es una práctica que usan muchas compañías. Como hay cada vez más instituciones públicas que abren sus datos, es también muy útil para periodistas que saben escribir código.
Los extractores de datos también se llaman “bots” o robots, y con ellos es posible extraer muchos datos.
El autor da el ejemplo de su historia en que compara los precios del alcohol en Ontario y Quebec.
Es relevante discutir la ética de la extracción de datos. Muchos manuales de ética periodística aún no lo incluyen, por lo que el autor entrevistó a varios periodistas de datos para responder ciertas preguntas.

Datos públicos, o no?

La regla es que si una institución pública libera datos, estos deben hacerse públicos automáticamente. Sin embargo, estas instituciones también guardan en sus servidores datos de ciudadanos.
William Wolfe-Wylie, desarrollador de CBC y profesor de periodismo en Centennial College y Munk School en la Universidad de Toronto, explica que esta información normalmente está escondida para no violar leyes de privacidad.
La diferencia entre el hackeo y el scraping - o extracción de datos-, es el respeto a las leyes. Los reporteros no deberían cazar data escondida, sino enfocarse en la que está de acceso a cualquier ciudadano y siempre leer los términos y condiciones del uso de la información.
Hay que chequear los archivos robots.txt del sitio, que establecen qué se puede extraer y qué no.

Identificarse, o no?

El periodista siempre debe identificarse como tal antes de hacer preguntas. Pero, ¿qué ocurre en el caso de un robot?
Glen McGregor, reportero de asuntos nacionales para el Ottawa Citizen, piensa que se debe seguir la misma regla. Dice que para evitar que quien maneja el sitio piense que está siendo hackeado, en el encabezado http siempre pone su nombre, su número de teléfono y un mensaje diciendo quién es y que pueden llamarlo por dudas o problemas.
No todos piensan lo mismo. Philippe Gohier, editor web jefe de L'Actualité, hace todo lo posible para que no lo identifiquen: a veces usa proxys, cambia su IP y sus encabezados para ocultar que es un robot; así respeta las reglas pero se mantiene anónimo.
Para el autor del artículo, no identificarse al extraer datos equivale a usar cámaras espías, a infiltrarse al reportear, o mentir en la identidad.
El Código de Ética FPJQ tiene las siguientes reglas para justificar procedimientos encubiertos, que deben ser excepcionales:
Si “la información buscada es de interés público definitivo; por ejemplo, en casos donde las acciones socialmente reprensibles deben ser expuestas”; “no puede obtenerse o verificarse por otros medios, u otros medios ya se han utilizado sin éxito”; y “el beneficio público es mayor que cualquier inconveniente para las personas”.
Además, añade que el público debe ser informado si se usa alguno de estos métodos para obtener la información.

Entonces, la mejor práctica es siempre identificar al robot. Pero si hay riesgo de que la institución esconda información, por ejemplo, es mejor ser discreto sobre su identidad. De todos modos, si hay miedo de que el robot sea bloqueado, fácilmente puede cambiar su dirección de IP y ese problema se soluciona.
Otros periodistas piensan que lo mejor es pedir los datos primero, y extraerlos si es que son negados. Lo bueno de eso, es que si se entregan los datos el reportero ahorra mucho tiempo.

Publica su código, o no?

La transparencia es clave para que los periodistas tengan credibilidad frente al público.
La mayoría de los reporteros es transparente sobre los datos en que se basan sus historias.
En el caso de los códigos, si hubiese un error en su escritura, los datos obtenidos pueden llevar a historias completamente erradas.
Si se usa código abierto, es obligatorio revelarlo para que otros lo mejoren, y para que pueda ser fiscalizado.
Cuando está escrito para hacer una historia periodística es más compleja la decisión, porque ese código da ventajas sobre la competencia. Por eso, el extractor de datos Roberto Rocha, piensa que no todos los códigos deben hacerse públicos. Él, como otros, tiene un GitHub donde publica algunos.
Jean-Hugues Roy piensa que los periodistas deben compartir el código, así como los científicos comparten sus metodologías, para ayudar a todos y todas, pero que hay excepciones. Ahora, por ejemplo, está trabajando en un código que le ha tomado mucho tiempo y no está seguro si lo publicará.
Glen McGregor no lo publica, pero lo comparte si alguien lo solicita.
Cuando un reportero tiene una fuente hace todo para protegerla. Esto para ganar su confianza, pero también para mantenerla para sí mismo. El extractor de datos es como una fuente periodística.
Se discute también si los códigos periodísticos serán patentados en el futuro.
Un detalle técnico relevante: “respetar la infraestructura web es, por supuesto, otra regla de oro de la extracción web. Siempre deje varios segundos entre sus solicitudes y no sobrecargue los servidores".