MÓDULO 4

SEGUIMIENTO A LAS POLÍTICAS PÚBLICAS
SOBRE DESARROLLO SOSTENIBLE

PERIODISMO DE DATOS
PARA EL DESARROLLO SOSTENIBLE

EN EL MUNDO DE LOS DATOS:
EXTRACCIÓN O SCRAPING

En el mundo técnico asociado a la utilización de datos es común escuchar el término *scraping* o "escrapear" que si bien no es una palabra que exista en el idioma español, su apropiación y uso se volvió generalizado con el pasar del tiempo. Para quien usa esta palabra, comúnmente se refiere al concepto de extraer información. Es decir, el *scraping* es la extracción de información desde una fuente, sitio o documento no "abierto".

Ya hicimos mucho hincapié en la necesidad de contar con bases de datos "limpias", lo que en muchos casos no se presenta dadas las formas en que los datos son construidos desde un inicio o la manera de entregar la información. Pero ¿qué pasa con la información que no está en un formato abierto, o incluso no está en una base de datos? ¿Se le puede dar uso? Para ambas preguntas la respuesta es el uso de la **técnica de extracción de datos.**

La extracción de datos es útil en un momento específico de todo proyecto de uso de datos, cuando se recopila información de todo tipo de fuentes. La extracción a temprana edad del proyecto permite recopilar datos de fuentes poco amigables e igualmente valiosas sin importar el formato, lo que enriquece la colección de datos para su análisis o interpretación posterior.

En otros casos, cuando la información es pública pero no necesariamente útil por la forma en que se entrega (fotografías, escaneos de documentos, documentos escritos a mano, etc.) la extracción de datos puede ayudar a simplificar un proceso de conversión o captura de información, aunque igualmente requiere de un trabajo de procesamiento y limpieza posterior para garantizar la integridad de lo datos.

¿CUÁNDO SE PUEDE HACER UNA EXTRACCIÓN DE DATOS?

La extracción de datos cobra sentido cuando al ser empleada permite agilizar un trabajo que de forma manual tomaría mucho tiempo completarlo. Sin embargo, también se debe tomar en cuenta el tipo de información que se pretende recopilar y la calidad final que se tendría sobre la información al concluir el proceso, ya que en ocasiones si bien se puede avanzar ágilmente en extraer un dato o documento, la calidad final no es la suficiente y se debe emplear mucho tiempo en limpiar y estandarizar datos. La suma del tiempo de ambos procesos probablemente se pudo emplear desde un inicio para capturar la información.

En situaciones como la anterior, es importante ponderar el tiempo de trabajo invertido versus el resultado final, de forma que se pueda tener claridad sobre la calidad de datos que se obtienen al final de cada proceso de extracción de datos; al hacerlo se logrará emplear de la forma más eficiente los recursos disponibles.

PDFS CON IMÁGENES

En algunos casos la información se entrega en archivos PDF pero en su interior se presentan imágenes ya sean fotografías o

escaneos de originales que contienen los datos necesarios. Para esas situaciones lo ideal es usar herramientas de reconocimiento de caracteres o también conocidos como OCR. Estos programas lo que hacen es usar un algoritmo que permite reconocer cada letra como un símbolo, de tal forma que al hacer una inspección profunda de los archivos pueden reconstruir en gran parte un archivo con sólo hacer una inspección del documento.

En gran medida, esta es una de las herramientas y técnicas más usadas para extraer datos por dos razones: 1) porque en muchos casos los archivos PDF contienen imágenes de los datos y 2) porque el software OCR que se emplea para lograr la extracción funciona con un grado de precisión bastante alto; lo que genera que el documento final sea de alto valor para el usuario.

PDF CON DATOS TABULADOS

Otro tipo de archivos a los cuales podemos enfrentarnos pueden ser los archivos PDF que en su interior contengan datos tabulados o tablas de información. En muchos casos este tipo de documentos se entregan así porque desde su origen el sistema que genera los datos permite exportar la información en formato PDF y, por tanto, no se hace un proceso para "liberar" los datos como se mencionó en el módulo 1 sobre calidad de los datos.

Ahora, para extraer datos tabulados se pueden usar un par de herramientas. La primera de ellas y muy popular es Tabula PDF. Esta herramienta permite reconocer la ubicación de los datos tabulares y procesarlos de tal forma que los ordena y acomoda en un patrón listo para su análisis. La herramienta es realmente sencilla para usar.

Un detalle importante con Tabula es que en archivos de gran escala el proceso puede ser un poco más lento de lo normal por el intensivo uso de recursos (procesador y memoria ram de la computadora) necesario para completar la tarea.

EXTRACCIÓN DESDE SITIOS WEB

Aunque invisibles, en algunos casos muchas páginas web están construidas de tal forma que en su interior contienen una serie de tablas en las cuales se organiza la información que se puede ver, sin embargo, las tablas no son visibles ya que no se les asigna un color que las distinga. La extracción de datos también puede ser usada para obtener datos de sitios web que contengan este tipo de tablas tan distintivas.

Extraer datos de este tipo de fuentes puede requerir cierto conocimiento técnico que le permita a los usuarios entender ciertas complejidades detrás del diseño y funcionamiento de una página web. Sin embargo, esto no limita que en las formas más simples de páginas web cualquier usuario pueda extraer información para una base de datos.

Existen diversos complementos para navegadores o para sistemas como hojas de cálculo que se pueden usar para extraer información con unos cuantos clics y sin mucho conocimiento técnico previo. Estos complementos en su mayoría son gratuitos y se obtienen en diversas "tiendas" de aplicaciones para navegadores.

Si está en México, Colombia o Costa Rica, pase por el ecosistema de datos de **www.datarepublica.org** y explore un poco el panorama de la disponibilidad de datos para cada Objetivo de Desarrollo Sostenible (ODS) en estos países.

DATAREPÚBLICA

Cepei
Knowledge factory for the world we deserve