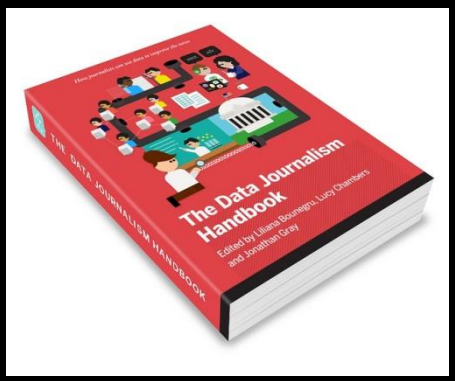


Periodismo de base de datos y transparencia

Equipo de Transparencia - DGSC

DIRECCIÓN GENERAL DE SERVICIO CIVIL
DGSC
Presidencia de la República
CostaRica

Cápsula Informativa
Equipo de Transparencia DGSC
No. 6
Agosto 2016

Periodismo de base de datos y transparencia

¿Qué es el Periodismo de Datos y Búsqueda de Datos?

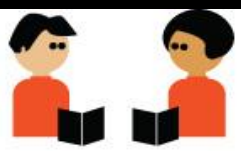
Por Sandra Crucianelli
Experta en *Deep Searching Web* y Bases de Datos. Especialista en Periodismo de Investigación y Periodismo de Precisión, con énfasis en fuentes digitales y Periodismo de datos.
Docente del Centro Knight Internacional para periodistas.

El periodismo de base de datos, o periodismo de datos como se lo conoce con más frecuencia, es una combinación entre el periodismo de investigación de siempre, con métodos del periodismo de profundidad, precisión, analítico y del "asistido por computadora".

En este tipo de periodismo se trabaja con un gran volumen de datos abiertos, muchos de los cuales provienen de formatos cerrados, por lo que en estos casos hay que hacer una tarea previa de apertura de datos.

Los 3 sellos del Periodismo de Datos

Ofrece a la audiencia los documentos de respaldo sobre los que se realizó la investigación periodística, generalmente compartidos desde una plataforma externa.	El reportero explica sus métodos de investigación, de modo que si un lector u otro periodista quiere recorrer el mismo camino usando los mismos documentos, puede hacerlo. Esto le permitirá apoyar o refutar los resultados obtenidos.	Incluye una adecuada visualización de los datos, mediante cuadros, gráficos, infografías, mismos que se acompañan de textos cortos.

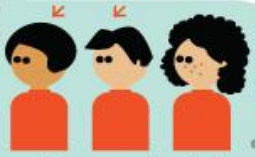


PERIODISMO DE DATOS
Manual

1 Introducción
¿Qué es el periodismo de datos?

2 En la redacción
La cuestión de fondo

Periodista Programador



Hacker se alquila

El armado del equipo

3 Estudios de casos
Ejemplos que inspiran

4 Obtener datos
Datos

Acceso a información pública

Portales de datos abiertos

Scraping (rascado)

Crowdsourcing (colaboración abierta)

Aspectos legales en los datos

5 Comprender los datos
Herramientas del oficio

Herramientas

Saber leer datos

6 Transmitir datos
Contar historias con datos

Presentar datos en notas

Aplicaciones con datos

Visualización de datos

Periodismo de datos

Audiencia

Participación de la comunidad

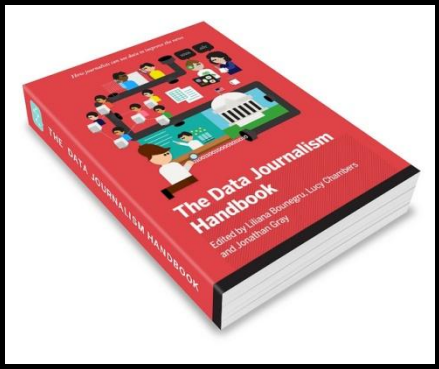
Los productos del periodismo de datos

- Artículos basados en datos.
- Visualizaciones Interactivas.
- Conjuntos de datos abiertos o Datasets.
- Aplicaciones de Noticias o News Apps .
- Blogs de datos
- Canales de datos o Sección de Datos.
- Catálogos de datos .

Mediante estos productos informativos, la prensa está dando una nueva dimensión a la transparencia, ya que se extraen datos y se reinterpretan desde una nueva visión.

Las versiones oficiales están siendo escrutadas.

Extracción de datos (web scraping)



Una de las técnicas que actualmente utilizan con mayor frecuencia periodistas e investigadores es la extracción de datos que han sido publicados en formatos cerrados. El ejemplo más común son los archivos PDF.

Mediante la técnica del *web scraping* se extraen datos escondidos en un documento, como páginas web y PDF, y los hace útiles para usarlos después. Para ello hace uso de diversos software, tales como: Zamzar.com, que es gratuito y no requiere suscripción; Free Ocr, eficaz con textos pequeños, sin sellos ni firmas manuscritas; Document Cloud: https://www.documentcloud.org/ ; tabula, http://tabula.nerdpower.org/.

Cómo armar un set de datos

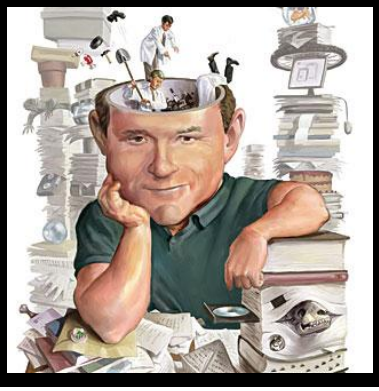
Las formas más comunes de gestionar y alojar sets de datos son a través de planillas de Excel o mediante un formato denominado CSV, abierto y reutilizable, en el que las columnas están separadas por comas.

Cuando se descarga un fichero de datos en CSV debe seguirse un procedimiento para armar de nuevo los datos en columnas, en una hoja de cálculo y realizar el análisis de los datos.

Lista de herramientas de scraping de datos en:

https://docs.google.com/spreadsheets/d/1RNuyJbt2Mz9KlZ_HWpfAg-u4vJNbUJcD4fsWiHMXByw/edit#gid=1060952292

Minería de datos



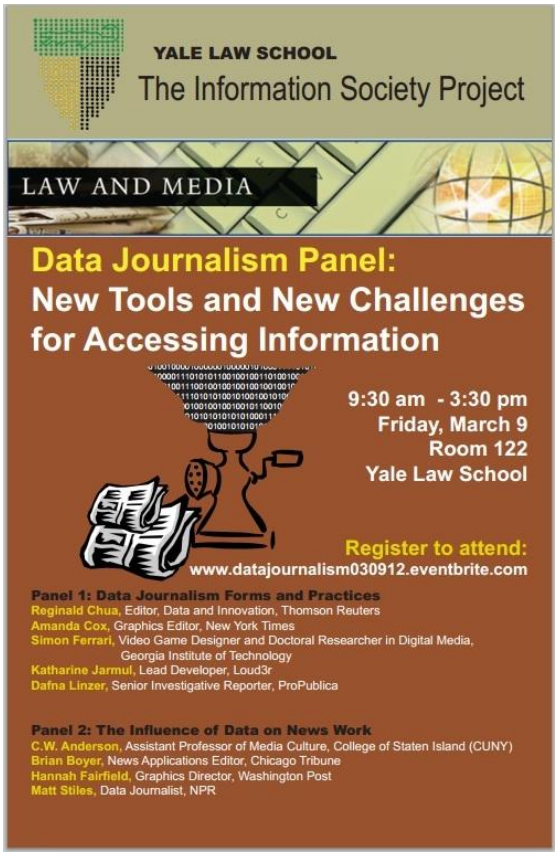
La minería de datos o exploración de datos es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.

Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos.

El objetivo general del proceso es extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.

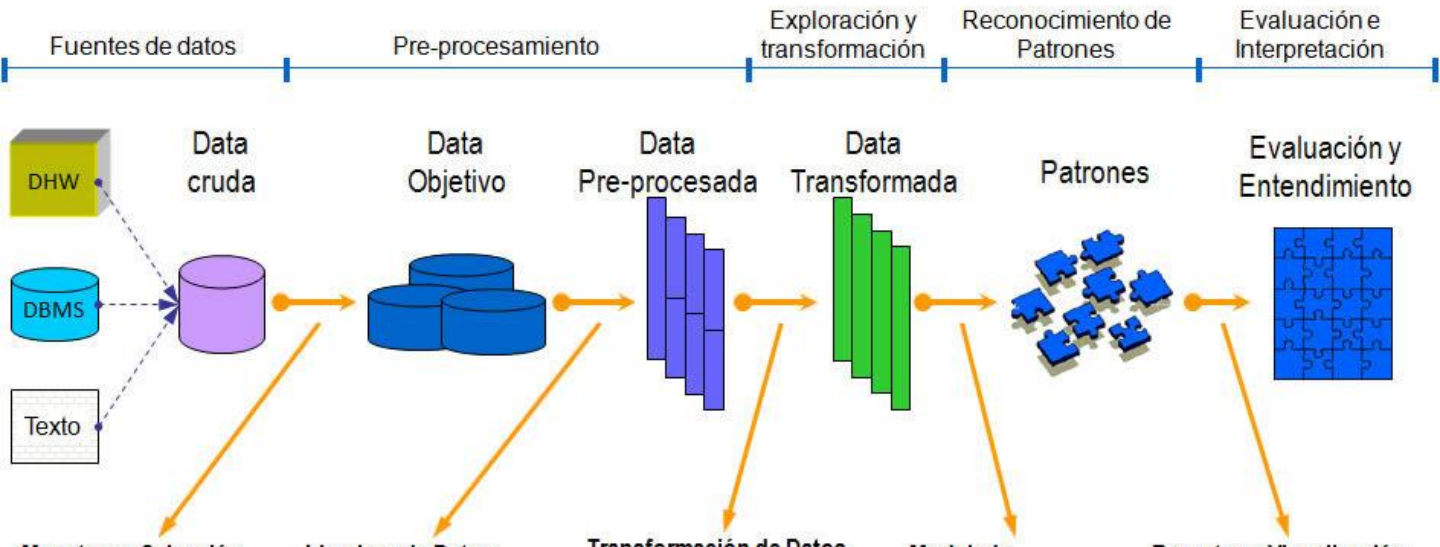
Además de la etapa de análisis en bruto, que involucra aspectos de bases de datos y de gestión de datos, de procesamiento de datos, del modelo y de las consideraciones de inferencia, de métricas de Intereses, de consideraciones de la Teoría de la complejidad computacional, de post-procesamiento de las estructuras descubiertas, de la visualización y de la actualización en línea.

La tarea de la minería de datos real es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos.



Minería de datos

El Proceso de la Minería de Datos



Muestreo y Selección
- Muestreo
- Selección

Limpieza de Datos
- Datos que no existen
- Datos no clasificados
- Identificación de extremos
- Eliminación de Ruido

Transformación de Datos
- Reducción de Dimensionalidad
- Creación de Características
- Normalización de Datos
- Variables Correlacionadas
- Discretización

Modelado
- Clasificación (P)
- Regresión (P)
- Agrupamiento (D)
- Asociación (D)
- Secuenciación (D)

(P) Predictivo
(D) Descripitivo

Reportes y Visualización

Simple
Complejo

Bibliografía
Data Mining: Practical Machine Learning Tools and Techniques. Ian H. Witten, Eibe Frank. Morgan Kaufmann
Introducción a la minería de datos. J. Hernández, J. Ramírez

Felipe de Jesus Núñez Cárdenas
Profesor Tiempo Completo
Escuela Superior Huejutla
Universidad Autónoma del Estado de Hidalgo

¿Cómo buscar información en la web profunda?

- La periodista Sandra Crucianelli señala las siguientes formas de obtener datos:
- Rastreo de la web, conocido como *searching*.
- Rastreo en la **"web superficial"** donde es posible encontrar los resultados más comunes que devuelven los buscadores, como páginas de sitios comerciales o aquellos con alto tráfico.

- Rastreo en la **"web profunda"** o *deep web*, espacio donde se incluye información que no es indexada por los motores de búsqueda como Google, Bing, etc.
- En ella es posible encontrar documentos que no han sido almacenados bajo estructura HTML, como archivos para abrir o descargar en diferentes formatos: PDF, Excel, PPT (Power Point), incluso Flash y las extensiones que utiliza Google Earth.