Consideraciones éticas en la obtención de información en Internet, big data y analítica web

JOSÉ ANTONIO FRÍAS
UNIVERSIDAD DE SALAMANCA

Índice

- ¿Qué es el Big Data?
- Datificación
- El valor de los datos
- Ejemplos de Big Data
- Desafíos del Big Data
- Consideraciones éticas

¿Qué es Big Data?

"Volumen masivo de datos, tanto estructurados como no-estructurados, los cuales son demasiado grandes y difíciles de procesar con las bases de datos y el software tradicionales" (ONU, 2012)

Definición

- Posiblemente la próxima tecnología revolucionaria en Tecnologías de la Información

- Apareció en la primera década del siglo XXI

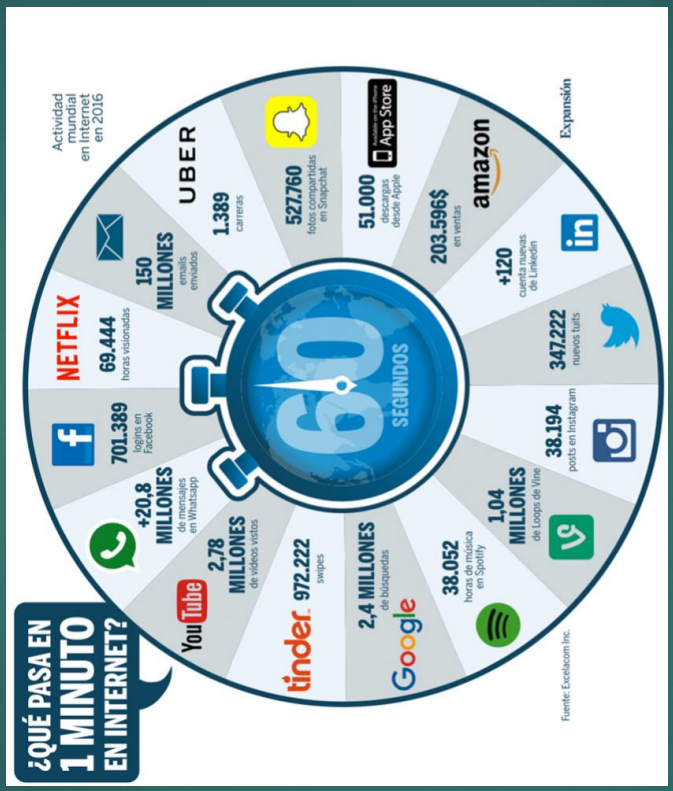
- Big Data tiene como objetivo resolver problemas antiguos y nuevos de manera más eficiente pero aplicado a extensos volúmenes de datos no usados hasta el momento

- Genera valor a un negocio a partir del almacenamiento y procesamiento de cantidades muy grandes de información digital que no puede ser analizada con técnicas tradicionales de computación

Almacenamiento actual

- Las capacidades de almacenamiento de datos de los sistemas de información actuales son enormes
- Actualmente existen almacenados > 2.7 Zetabytes (1 Zetabyte = 1 Trillón de gigabytes), se esperan 35 Zetabytes para 2020
- En 2012 la información digital alcanzó a nivel mundial 2.837 exabytes (miles de millones de gigabytes)
- Puestos en DVDs, la torre sería de 400.000 Kms, más que la distancia de la Tierra a la Luna
- Google procesa más de 24 Petabytes/día, información equivalente a varios miles de veces la biblioteca del congreso de USA
- En 2007 solo el 7% de la información estaba en medios analógicos (libros, revistas, fotografías en papel, etc.)

Escuela Graduada de Ciencias y Tecnologías de la Información (EGCTI-UPR)



Algunas experiencias gubernamentales a nivel internacional

Corea del Sur: “Plan Maestro de Big Data para la Implementación de una Nación Inteligente” (2013), del gobierno coreano.

Estados Unidos: “Iniciativa de I+D en Big Data” (2012), propuesta de la administración Obama, dirigido por la Oficina para la Ciencia y la Tecnología de la Casa Blanca.

Japón: Dentro de la primera estrategia de crecimiento del Japón del gobiernode Shinzo Abe (“Desatar el poder del sector privado hasta su máxima extensión”), se encuentra un plan básico para aprovechar Big Data” (Mayo 2012).

Comisión Estadística de Naciones Unidas: Seminario de Asuntos Emergentes en la 44° Sesión de la Comisión: Big Data para la Política, el Desarrollo y las Estadísticas Oficiales

¿Qué se puede hacer con todos estos datos?

- Imposible analizar con las técnicas tradiciones de BBDD (no da tiempo)
- Imposible almacenarlas siguiendo el modelo clásico de BBDD
- Recurrir a técnicas estadísticas de análisis de datos
- Sabremos *qué* pasa, pero no *por qué*. Descubrir nueva información que no está explicita a primera vista
- Pasamos de análisis a nivel local a análisis a nivel global

Big Data no está orientada a enseñar a los ordenadores a "pensar" como humanos, sino a aplicar estadísticas a los datos para obtener probabilidades de que ocurran ciertos sucesos

Características del Big Data

- N = ALL
  - Fin del muestreo
  - Todo el conjunto de datos es válido, no se descartan casos
  - Se puede eliminar el problema del sesgo
- La inexactitud de los datos ya no es un problema
  - El error de la muestra se minimiza, y por tanto podemos asumir datos menos exactos
  - Dicho de otra forma, cuando dependemos de una muestra, queremos que los datos sean exactos
- Obtenemos el qué, pero no el por qué
  - Las técnicas de Big Data no explican la causalidad (correlaciones)

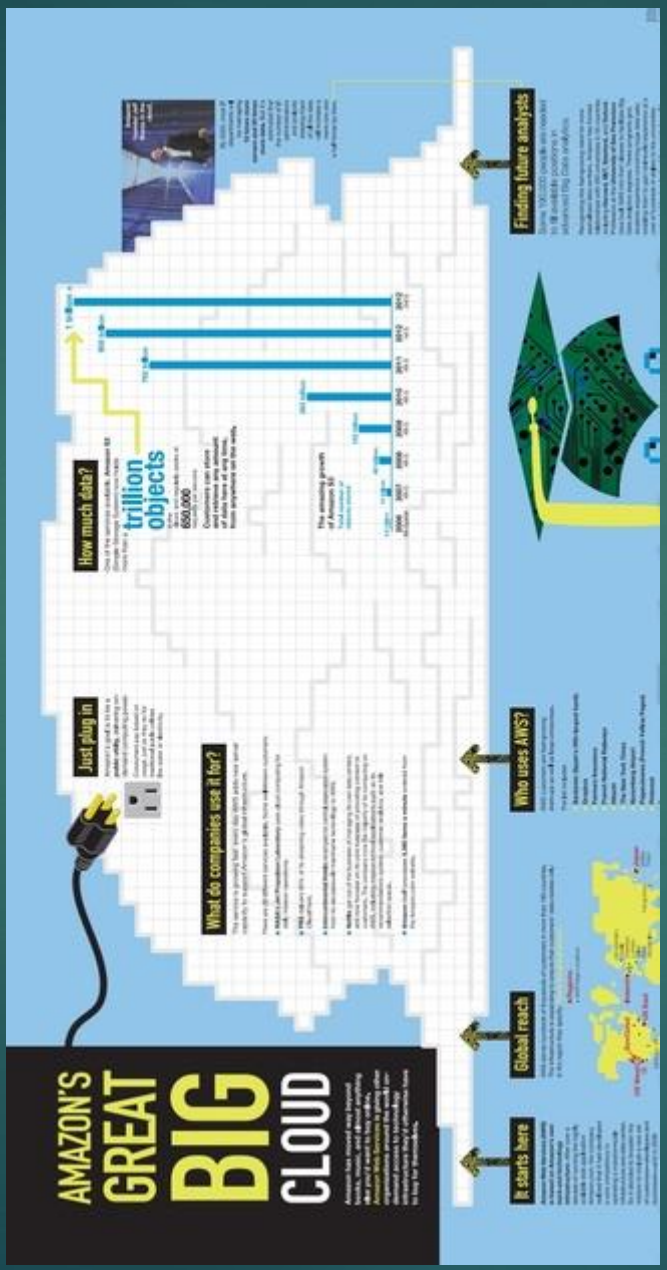
Las tres ‘v’ del Big Data

- **Volumen**: Se considera un volumen grande a partir de Petabytes (1.000.000 GB)
- **Velocidad**: Frecuencia a la que se genera los datos / Tiempo de análisis de los datos
- **Variedad**: Datos estructurados, semi-estructurados, no estructurados
- ¿Otras Vs?: ¿Veracidad? ¿Valor?

Tipos de datos

- Estructurados
  - Bases de datos tradicionales
  - Encuestas, censos
- Semi-estructurados
  - XML, RDF
  - Grafos
- No estructurados
  - Texto
  - Imágenes, audio

Escuela Graduada de Ciencias y Tecnologías de la Información (EGCTI-UPR)



"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Datificación

- Es el proceso de monitorizar una actividad que era "invisible" y convertirla en datos digitales.
- Estos datos eran ignorados porque no se disponía de tecnología para medirlos
- Ahora, toda acción, dato, preferencia, etc. está sujeta a ser medida y almacenada
  - Localización de una persona
  - Vibración del motor de un coche
  - Recorrido de una camilla en el hospital

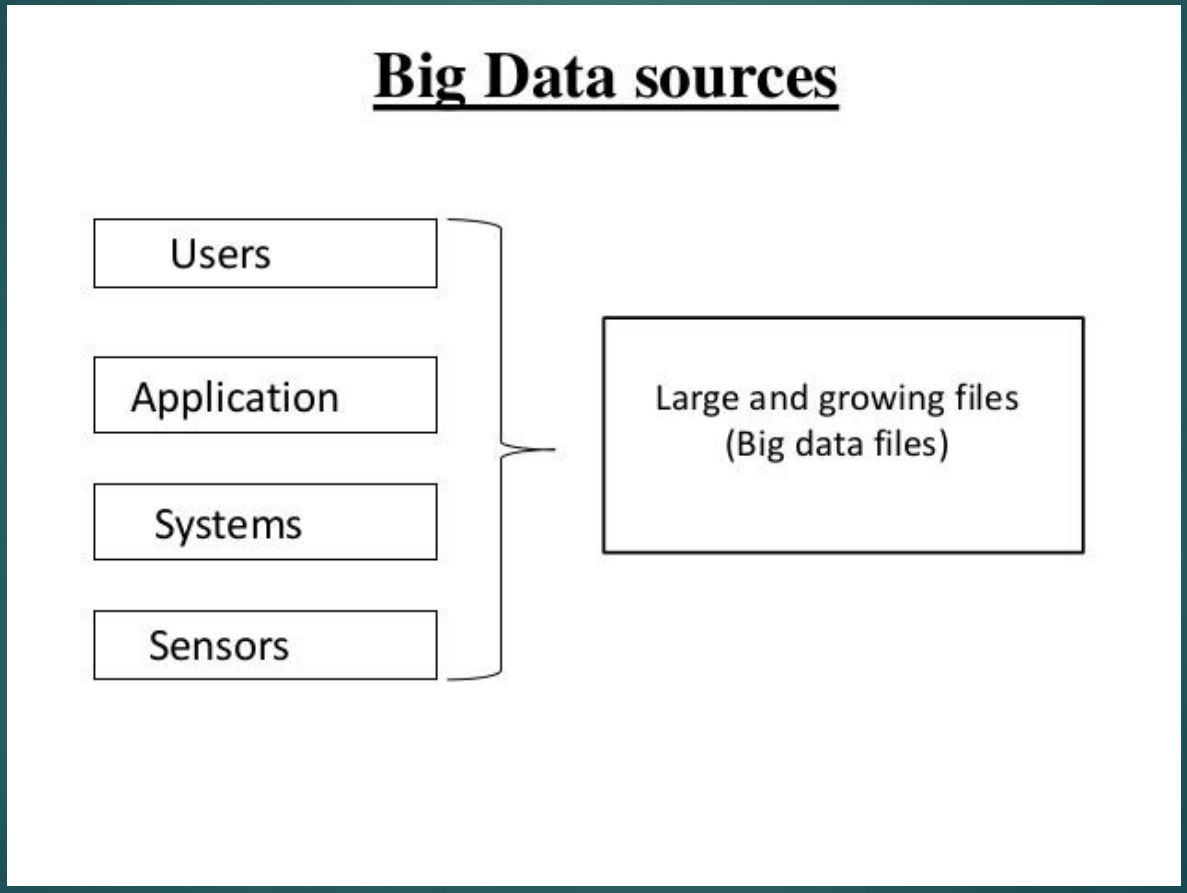
Datificación

- Con la datificación, los datos se pueden cuantificar de manera que pueden ser tabulados y analizados
- La digitalización de todo posible dato en varios formatos y guardarlos en formato estructurado permite el análisis en Big Data
- Dicho análisis permite detectar patrones de comportamiento, por ejemplo en las preferencias de clientes

Fuentes de datos

- Usuarios
- Aplicaciones
- Sistemas
- Bases de datos tradicionales
- Sensores (en vehículos, edificios, almacenes, productos, etc.)

Fuentes de datos



Usos

- La información recogida se puede utilizar posteriormente para fines con los que no se contaba en un principio o que no tienen nada que ver con el uso principal de esos datos
  - Postura de una persona al sentarse: Antirrobo de coche
  - Facebook: Concesión de créditos
  - Móvil: Detección temprana de Parkinson
  - Lectura de un libro: Detectar pasajes "difíciles de leer" para autores

Ejemplo: Culturomics

- Google dispone de un servicio de digitalización de libros
- Pensado para preservar libros antiguos
- Dispone de más de 95 billones de líneas de texto en múltiples idiomas
- Nuevo servicio: **Culturomics**, estudio del comportamiento humano y tendencias culturales a partir del análisis cuantitativo de texto digitalizado
  - https://books.google.com/ngrams/

El valor de los datos

- El uso intensivo de los datos ha pasado a ser el petróleo de muchas compañías
- El nuevo enfoque es almacenar cualquier tipo de dato, por irrelevante que pueda parecer, para su posterior análisis
  - Clicks de ratón en la página web de mi negocio
  - Vibración del motor del coche
  - Movimiento del acelerómetro del smartphone
- Permite crear modelos para responder preguntas complejas, mostrar percepciones contraintuitivas y aprender resultados únicos

Escuela Graduada de Ciencias y Tecnologías de la Información (EGCTI-UPR)

Roles en Big Data

1. **Almacenadores de datos**

- No utilizan los datos más allá de su uso primario definido en su negocio
- Pueden licenciar los datos a otras compañías para obtener beneficio económico
- Se convierten en proveedores de datos
- Ejemplos:
  - Twitter
  - Compañías telefónicas
  - Compañías de tarjetas de crédito

Roles en Big Data

2. Analistas

- No suelen disponer de los datos primarios, pueden adquirirlos
- No poseen la idea del valor añadido de los datos (no saben para qué pueden servir)
- Realizan análisis estadísticos y de minería de datos sobre enormes volúmenes de datos
- Ejemplos:
  - Consultoras
  - Empresas tecnológicas
  - Departamentos de estadística

Roles en Big Data

3. Big Data puro

- No suelen disponer de los datos primarios ni tienen las herramientas para realizar su análisis
- Poseen una idea novedosa sobre un valor añadido de los datos no extraído hasta el momento
- Ejemplos:
  - Jetpack
  - FlightCaster

Data Scientist

- Se ha creado un nuevo perfil llamado "científico de datos" altamente demandado en este sector
  - Un tercio administrado de base de datos, un tercio estadístico, un tercio gestor
  - Incluido en los 25 mejores trabajos en USA en 2015 (Glassdor report)

Data Scientist

Escuela Graduada de Ciencias y Tecnologías de la Información (EGCTI-UPR)

Data Scientist

Ejemplos de Big Data y casos de éxito

Escuela Graduada de Ciencias y Tecnologías de la Información (EGCTI-UPR)

Google y la predicción de la gripe H1N1 (2009)

- Más rápido que el propio sistema de salud de USA (llegaba con dos semanas de retraso)
- Correlación de términos más buscados usando previas temporadas de gripe
- Resultado: Encontrados los términos de búsqueda más relacionados con la gripe H1N1

Farecast.com

- Comprada por Microsoft en 2008 y hoy desaparecida
- Permitía predecir los cambios de precios de billetes de avión
- Contenía una base de datos con rutas y los precios de cada vuelo según día, climatología, etc
- Kayak, Travel MSN son sucesoras basadas en esta idea

Traductor de Google / Traductor IBM

- Traductor IBM cuenta con varios diccionarios (millones de palabras) y complicados algoritmos de traducción desde los años 60
- Traductor de Google cuenta con millones y millones de páginas Web en más de cincuentas idiomas y procesos simples de comparación de misma página en distintos idiomas
- ¿Cuál funciona mejor?

Amazon.com

- Inicialmente contaba con críticos (humanos) para aconsejar a sus clientes sobre qué libros comprar
- Aunque funcionaba bien, la siguiente idea fue personalizar las recomendaciones en base a anteriores compras, libros por los que el cliente navega en la página, libros similares, etc.
- Las ventas se dispararon mejorando de manera drástica la técnica de las críticas
- Actualmente ofrece conjunto de datos públicos para Big Data

Walmart

- Cadena de supermecados en USA
- Registra las compras de todos sus clientes y relaciona los productos que se suelen comprar juntos
- El sistema aconseja ofertas a clientes basadas en sus compras anteriores
- También indica a los responsables de los supermecados qué tipo de ofertas conjuntas deben hacer para promocionar artículos poco vendidos junto a best-sellers o eventos tales como huracanes (ocurre con las tartas de fresa)

Control de explosiones de alcantarillas en NY

- Cada año varios cientos de alcantarillas eléctricas explotan en NY, a veces expulsando la tapa de 135 Kgs por los aires
- Con Edison, la empresa responsable, realizaba inspecciones periódicas, pero no podía abarcar toda la red de alcantarillado. La tasa de aciertos estaba debajo del 5%
- Se utilizaron los datos (¡¡desde 1880!!) del tipo de cable, años en funcionamiento, explosiones previas, etc., para intentar predecir la siguiente explosión
- La predicción funcionó de manera que acertaron el 45% de las 100 próximas explosiones

Claves del Big Data

- **Integración**: Una única plataforma para manejar los datos
- **Análisis**: Preprocesamiento de datos + análisis estadístico
- **Visualización**: Cómo mostrar los datos a todo tipo de usuarios
- **Desarrollo**: Necesidad de herramientas sofisticadas
- **Seguridad y gobernanza**: Políticas de privacidad de datos

¿Qué hace inteligente a un sistema de Big Data?

- **Análisis**: Manejar grandes cantidades de datos, no sólo los directamente relacionados con el problema, sino también cualquier otra dato que pueda influir en la decisión
- **Instrumentos**: Para recoger los datos debe existir un conjunto de instrumentos capaces de medir la información deseada
- **Interconexión:** Debe existir una infraestructura preparada para recibir y almacenar cualquier tipo de datos recogido y procesarlos de manera eficiente

Gobernanza de datos (Data governance)

- ¿Podemos confiar en la(s) fuente(s) proveedora(s) de datos?
- Las empresas deben asegurar la confianza de sus fuentes de datos y protegerlas
- Debe existir una gobernanza de los datos durante el ciclo de vida de la información
- La gobernanza de datos debe estar integrada en la plataforma de manejo de los datos

Beneficios y riesgos

- Big Data no es la solución a todos los problemas.
- Las predicciones realizadas no son siempre correctas. Riesgo de caer en la "dictadura de los datos"
  - No mirar nada más allá de la información que nos den los datos
- ¿Perdemos privacidad?
  - ¿Sabemos los usos secundarios que le pueden dar a nuestro datos que los damos para otro fin primario distinto de su futuro uso?
  - ¿Puede una compañía de seguros mirar nuestros datos para saber si somos "asegurables"?
  - ¿Minority Report?
- Será necesaria una legislación sobre el uso de los datos y el "derecho al olvido"

Herramientas Big Data

- Plataformas Big Data:
  - Hadoop
  - HortonWorks
  - Cloudera
- Data Warehouse:
  - InfiniDB
  - Oracle
- Minería de Datos:
  - Weka
  - Rapidminer

Herramientas Big Data

- Bases de datos NoSQL:
  - MongoDB,
  - Cassandra,
  - Redis, etc.
- Lenguajes de programación
  - CUDA
  - OpenGL
- Generadores de análisis estadísticos:
  - SPSS,
  - R,
  - Talend Open Studio,
  - Skytree server, etc.

Herramientas Big Data

- Preprocesamiento de datos
  - Expresso
  - Curl
  - Spark MLP

Desafíos para las bibliotecas

GESTIÓN DE LAS BIBLIOTECAS BASADA EN EL….
¿DATO?

Desafíos para las bibliotecas

Aparte del presupuesto, claro

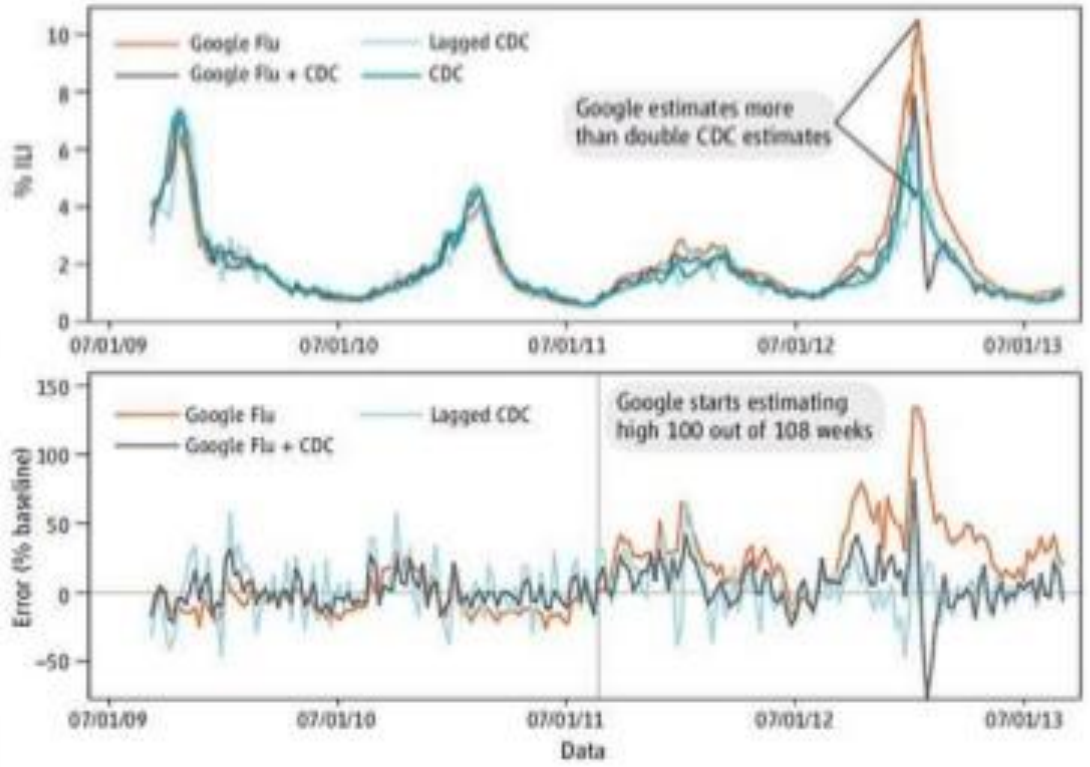
- **Privacidad de los datos personales**

http://www.dilbert.com

Desafíos para las bibliotecas

Aparte del presupuesto, claro

- **Confianza ciega en los datos (caso Google Flu)**



http://scholar.harvard.edu/files/gking/files/0314policyforumff.pdf

Oportunidades para las bibliotecas

GESTIÓN DE LAS BIBLIOTECAS BASADA EN EL….
¿DATO?

Roles de la biblioteca en big data

- Emisora de datos
- Garante de los datos
- Integradora de datos
- Analista/explotadora de datos

Escuela Graduada de Ciencias y Tecnologías de la Información (EGCTI-UPR)

La biblioteca, emisora de datos

Escuela Graduada de Ciencias y Tecnologías de la Información (EGCTI-UPR)

La biblioteca, garante de los datos

Biblioteca Nacional de España: dominio ".es"

- Abdicación de Juan Carlos I
- Proclamación de Felipe VI
- Muerte de Adolfo Suárez
- Muerte de Ana María Matute
- Muerte de Gabriel García Márquez
- Atentado de Charlie Hebdo

La biblioteca, garante de los datos

Algunas iniciativas:

- re3data.org: Registro global de repositorios de datos de investigación.
- DataHub: Trabajan en estándares para compartir datos de todo tipo.
- SPARC Europe

La biblioteca, garante de los datos

Algunas guías:

- Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020.
- Research Data Management (Duke University Libraries).

La biblioteca, integradora de datos

- Data sharing, Linked Open Data…
- Web scale discovery tools

Escuela Graduada de Ciencias y Tecnologías de la Información (EGCTI-UPR)

La biblioteca, analista / explotadora de datos

- Integración de datos
- Organización del conocimiento (control del vocabulario, ayuda en la recuperación de información…)

Consideraciones éticas para la obtención de datos de carácter personal en Internet

El marco de la acción ética en Internet (Estalella y Ardèvol)

- Las guías éticas como orientadoras de la práctica etnográfica en Internet.
- La ética en Internet.
- Dicotomía público/privado: ¿dónde está la frontera?
- Los dilemas éticos pueden tener varias soluciones
- Más allá de la tecnología contextual: la ética dialógica

Las guías éticas como orientadoras de la práctica etnográfica en Internet

- **Imperativo categórico** de los códigos deontológicos: la investigación debe **evitar dañar** a quienes forman parte del estudio, así como respetar su **seguridad** y **privacidad**.
- Mecanismos más utilizados:
  - Anonimato de los datos personales obtenidos.
  - Consentimiento informado.

Guías éticas para la investigación en Internet

- *Ethical Decision-making and Internet Research Recommendations from the AoIR Ethics Working Committee* (2002).
  - Dificultad de aplicar los principios éticos de la investigación al ámbito digital:
    - ¿qué significa mantener el anonimato de personas que charlan usando nombres de usuario en un chat de acceso público?, ¿qué interacciones son públicas en internet y no necesitan consentimiento informado?, ¿aquellas que son simplemente de "acceso público"?, un foro con clave de acceso, ¿es público?, ¿qué tipo de registro puede ser realizado sin necesidad de solicitar consentimiento?

Dicotomía público/privado: ¿dónde está la frontera?

- Algunos investigadores consideran que público en Internet son aquellas interacciones cuyo "acceso" es público.
  - "Cualquier persona que usa sistemas de comunicación disponibles públicamente en internet debe estar al corriente de que esos sistemas son, en su constitución y por definición, mecanismos para el almacenamiento, transmisión, y recuperación de comentarios. Que algunos participantes tengan cierta expectativa de privacidad, es algo erróneo" (WALTHER 2002).

Dicotomía público/privado: ¿dónde está la frontera?

- Otros investigadores han aportado evidencias de que las "expectativas de privacidad" de las personas que interaccionan a través de Internet no coinciden a menudo con las de un observador externo que no forma parte del colectivo.
- La experiencia que tienen los miembros de ese colectivo es de una relativa privacidad. Las normas de privacidad se desarrollan dentro de los colectivos, y no derivan únicamente de la configuración de la tecnología.

Dicotomía público/privado: algunas consideraciones

- La percepción de lo público y lo privado puede variar según la posición de la persona observada.
- La tecnología o la arquitectura tecnológica no determina el carácter privado o público de un espacio de interacción, sino que depende del sentido que le atribuye a esas interacciones cada colectivo.
- Lo público y lo privado no son categorías absolutas que podamos determinar "a priori" en las interacciones de Internet, son contextuales y dependen de la negociación que cada colectivo lleva a cabo.

Los dilemas éticos pueden tener varias soluciones

- "¿Cuál es la audiencia pretendida para una comunicación electrónica?, ¿te incluye a ti como investigador/a?" (FERRI 2000)
- "La violación de la privacidad se produce cuando las expectativas razonables de una persona son frustradas en lo relativo a las limitaciones del uso de la información personal" (ELGESEM 2002)
- LAS DECISIONES ÉTICAS DEBEN ORIENTARSE DESDE EL CONOCIMIENTO DE LA PERSPECTIVA DE LA OTRA PERSONA. Y, ANTE LA DUDA LA MEJOR OPCIÓN ES IR Y PREGUNTAR SIEMPRE QUE SEA POSIBLE.

Más allá de la tecnología contextual: la ética dialógica

- Es la expectativa de privacidad que tienen las personas la que debe servir de referente en nuestras decisiones.
- No es posible decidir si es público o privado considerando únicamente si se exige contraseña para acceder, o suponer que es anónimo porque no figura el nombre legal de la persona y sólo aparece un pseudónimo.
- Asumir esta posición nos obliga a que nuestras valoraciones éticas se orienten desde una actitud reflexiva que puede variar sobre la marcha, en función de cada nuevo contexto de investigación, y que evoluciona al contrastar nuestras percepciones con las de nuestros corresponsales.

Ejes sobre los que debe asentarse el compromiso ético durante la investigación

- la amplitud del registro de datos de cada persona,
- la declaración de la presencia del investigador y
- la búsqueda de simetría y mutualidad con los corresponsales en el campo

Alternativas de los investigadores para resolver problemas éticos

- Comités de ética
- Iniciativas participativas: *Problematorio*

José Antonio Frías

Universidad de Salamanca

frias@usal.es

Escuela Graduada de Ciencias y Tecnologías de la Información (EGCTI-UPR)