Laboratorio de Datos
Calidad de Datos

Clase de Viviana Cotik con
modificaciones de Pablo Turjanski

2do Cuatrimestre 2025

Calidad de datos

Clarín.com » Edición Viernes 14.02.2003 » Economía » **Un banco debe pagar $ 120.000 por**

Ediciones anteriores

FIGURAR EN UN LISTADO DE INCUMPLIDORES ARRUINO UN NEGOCIO

Un banco debe pagar $ 120.000 por incluir mal a un cliente en Veraz

Daniel Gutman

El Banco Río lo incluyó en las listas negras de deudores de la Organización Veraz y solicitó al Banco Central que lo inhabilitara. Pero todo era un error, porque no había existido ningún in cumplimiento. El cliente hizo juicio y obtuvo una sentencia de Cámara a su favor. Hasta ahí, un caso igual a muchos otros que ha habido en los últimos años. Lo novedoso es que la Cámara en lo Comercial acaba de establecer la que seguramente sea la indemnización más alta en este tipo de casos: el Banco Río deberá pagarle 120.000 pesos a su ex cliente.

A esa cifra deberán sumársele **los intereses** a la tasa activa del Banco Nación desde la fecha de inhabilitación en los registros del Central, que es mayo de 1996, lo que **llevaría la indemnización a más de medio millón de pesos**, según los abogados del demandante.

La importancia de la indemnización —según se explicó en el fallo— tiene que ver con que el damnificado **es un empresario que estaba en pleno proceso de ampliación de sus negocios.**

El hombre, dueño de una confitería, estaba construyendo un edificio en la avenida Cruz en el cual **pensaba instalar una concesionaria de autos**, además de una confitería y salón de fiestas en la planta alta. Sin embargo, en mayo de 1996 quedó sin posibilidad de obtener crédito y operar con cheques, por lo que **la obra y sus proyectos quedaron inconclusos.**

Así, la Sala B de la Cámara —en un voto de la jueza María de Díaz Cordero, al que adhirió Enrique Butty— aplicó el concepto de "pérdida de chance". Es decir, el Río deberá indemnizar al empresario porque **lo privó de una oportunidad de ganar dinero.**

Las pruebas presentadas y la trayectoria de Eloy Domínguez Alvarez convencieron a la jueza de que él "tenía intención de culminar con la construcción del edificio y ampliar sus negocios" y de que lo hubiera hecho "de no haber existido la arbitraria y errónea decisión adoptada por la entidad bancaria demandada".

Caso

1. Leer el artículo
2. En grupos de 3 integrantes …
   a. Describir el problema
   b. ¿Cuál es la causa del problema?
   c. ¿Quiénes se benefician al contar con datos de calidad?

Fuente: https://www.clarin.com/economia/banco-debe-pagar-120000-incluir-mal-cliente-veraz_0_rJ4juMlAFg.amp.html

Ejemplos de problemas

- Francisco L. Luján
  Anchorena 2236
  1425, CABA
  6082-2428

  Pancho Lujan
  Tomás Manuel de Anchorena 2236 5 B
  C1425FFG, Ciudad Autónoma de Buenos Aires
  4772-1656

- 30% de cartas rechazadas por problemas en la dirección
- Falta de conocimiento de a quiénes una compañía telefónica los está dejando sin teléfono sin aviso previo.

¿A quién le interesa contar con datos de buena calidad?

- Al investigador
- En la compañía
  - A quienes hacen uso de la información que proveen los datos
  - A los desarrolladores de los sistemas
  - A cualquier usuario de los datos
  - A sus clientes

¿Por qué es crítica la calidad de datos?

- El valor creciente de la información
    - datos vs información
- La información del sistema es “la que vale”

Temas

- Calidad de datos
- Algunas definiciones
- Ley de protección de datos personales
- Open Data (Datos abiertos)

Temas

- **Calidad de datos**
  - Introducción
  - **Definición**
  - (Ejercicio e Ingeniería reversa)
  - Problemas habituales
  - Causas
  - Atributos de calidad
  - Diagnóstico
  - Corrección y prevención
- Algunas Definiciones
- Ley de protección de datos personales
- Open Data / Datos abiertos

Calidad de Datos

¿Cuál es la definición de Calidad de Datos? (discutir 3 min.)

Calidad de datos (DQ-Data Quality-)

Podemos tener los mejores algoritmos, pero…

Garbage in – garbage out

Definición:

- “Un dato o conjunto de datos X tiene mayor calidad que un dato o conjunto de datos Y si X satisface las necesidades del usuario mejor que Y” [Redman, 1996]
- “Satisfacer de manera consistente las expectativas de los usuarios” [English, 1999]

Definición subjetiva

Calidad de Datos

¿Cuáles son las consecuencias de contar con datos de mala calidad?

(discutir 3 min.)

Calidad de datos

Gráfico tomado de AIT solutions

Consecuencias:

- descreimiento
- insatisfacción de clientes
- costos innecesarios
- impacto en toma de decisiones
- ...

Temas

- **Calidad de datos**
  - Introducción
  - Definición
  - **(Ejercicio e Ingeniería reversa)**
  - Problemas habituales
  - Causas
  - Atributos de calidad
  - Diagnóstico
  - Corrección y prevención
- Algunas Definiciones
- Ley de protección de datos personales
- Open Data / Datos abiertos

Trabajo en equipo

Ejercicio

- ✓ Conformar grupos de 3 integrantes
- ✓ Descargar del campus el registro de Datos de Dengue y Zika de 2020 correspondiente al Registro del Sistema Nacional de Vigilancia de la Salud 2.0
- ✓ Responder las siguiente preguntas:
  1. ¿Detectan problemas de calidad de datos? ¿Cuáles? ¿Pueden caracterizarlos?
  2. ¿Piensan que los datos provistos provienen de una única tabla de una BD relacional?

Problemas detectados

- Miramos en una base de datos. Algunas Alternativas: MySQL, POstgres, SQLIte.
  - BD vs SQL embebido.

Algunos problemas de DQ detectados

- datos nulos
  - departamento_nombre, provincia_nombre, provincia_id
- columnas intercambiadas
  - grupo Edad id vs grupo edad
  - envento_Nombre vs semanas_epidemiologicas
- sólo casos de dengue (no de zika)
- rangos erróneos en grupo etario
- aparentemente mismo departamento, pero escrito de manera distinta
  - Apostoles vs Apóstoles, etc.

Trabajo Grupal

- Conformar grupo de 3 estudiantes
- Discutir qué acciones tomarían (en este caso) para mejorar la calidad de los datos

Posible solución

Armar tablas normalizadas

1. Chequear y corregir problemas de calidad de datos
2. Crear nuevas tablas con id y descripciones y en la original dejar sólo los ids (ingeniería reversa)
3. Decidir qué hacer con registros que aparecen dos veces, pero con distinta cantidad de casos. Documentar la decisión.

Temas

- **Calidad de datos**
  - Introducción
  - Definición
  - (Ejercicio e Ingeniería reversa)
  - **Problemas habituales**
  - Causas
  - Atributos de calidad
  - Diagnóstico
  - Corrección y prevención
- Algunas Definiciones
- Ley de protección de datos personales
- Open Data / Datos abiertos

Algunos problemas habituales

- Valores no estandarizados
  - NETOFAGASTA
  - ANMTOFAGASTA
  - ANT0FAGASTA
  - ANTO9FAGASTA
  - ANTOAFAGASTA
  - ANTOFAAGASTA
- Valores imposibles o poco probables
  - Edad: 200 años
- Valores faltantes
  - Registros de personas con el campo e-mail vacío.
- Valores no actualizados

Algunos problemas habituales (parte 2)

- Ocurrencias duplicadas
- Falta de datos históricos
- Inconsistencia entre aplicaciones o en una misma aplicación
  - Datos de pacientes en dos servicios distintos de un hospital
  - Datos de pozos en dos aplicaciones distintas (perforación, producción)
- Información crítica que no es confiable
  - Hay postulantes a cursos, que han fallecido.

Temas

- **Calidad de datos**
  - Introducción
  - Definición
  - (Ejercicio e Ingeniería reversa)
  - Problemas habituales
  - **Causas**
  - Atributos de calidad
  - Diagnóstico
  - Corrección y prevención
- Algunas Definiciones
- Ley de protección de datos personales
- Open Data (Datos abiertos)

Trabajo Grupal

- Conformar grupo de 3 estudiantes
- Discutir cuáles pueden ser las causas por las que los datos tienen problemas de calidad

Calidad de datos

Posibles causas de problemas:

- procesos masivos que reparan un dato, pero no reconstruyen información relacionada
- misma información cargada en distintos sistemas
- valores predeterminados

Depende de:

- calidad de software
  - usabilidad
  - interfaz (obligatoriedad de carga)
  - seguridad
- definición de procesos asociados a los datos
- diseño de base de datos
- falta de capacitación
- ....

Posibles causas de problemas de DQ

Problemas asociados a:

- la instancia
- al modelo de datos
- a los procesos
- a errores de software

Asociados a la instancia

- datos que han cambiado en el mundo real, y que no fueron actualizados
- datos que provienen de distintas fuentes, deberían ser consistentes y sin embargo no lo son
- datos que no han sido almacenados con la precisión necesaria (por ejemplo, Y2K)

Posibles causas de problemas de DQ

Asociados al modelo de datos

- si se detecta que hay información que no está presente porque no hay forma de almacenarla -> el modelo de datos físico está incompleto
- el mundo que se quiere representar evolucionó y no se tradujeron los cambios al modelo -> pérdida de vigencia del modelo

Posibles causas de problemas de DQ

Asociados a los Procesos

- distintas personas cargan la misma información haciendo distintas asunciones
- se carga con una asunción y se usa con otras
- modificaciones manuales-por procesos
- gente que hace modificaciones pero no debería estar autorizada para hacerlas

Asociados a errores de software

- datos obligatorios que no se asumen como tales y por lo tanto no se cargan
- interfaces poco amigables

Posibles causas de problemas de DQ

El software de buena calidad no garantiza la calidad de la información

Se debe trabajar sobre:

- la instancia
- el modelo de datos
- los procesos que intervienen en la generación y modificación del dato
- la consistencia entre las diferentes fuentes de datos

Temas

- **Calidad de datos**
  - Introducción
  - Definición
  - (Ejercicio e Ingeniería reversa)
  - Problemas habituales
  - Causas
  - **Atributos de calidad**
  - Diagnóstico
  - Corrección y prevención
- Algunas Definiciones
- Ley de protección de datos personales
- Open Data (Datos abiertos)

Atributos de Calidad

¿Qué características deberían cumplir los datos para ser de calidad? (5 min.)

Calidad de Datos

¿Qué se requiere?

-datos completos

-datos oportunos (timeliness) y vigentes

-datos consistentes y correctos

-datos en cantidad adecuada

-datos disponibles/accesibles (ej. medicina), open data.

-datos seguros y privados (protección de datos personales)

Atributos de calidad

- completitud
  - están presentes todos los valores para representar la realidad
  - están presentes todas las instancias existentes en el mundo real
- relevancia
  - los datos son relevantes para representar la realidad
- vigencia
  - los datos se mantienen actualizados con la frecuencia adecuada
- disponibilidad
  - los datos están accesibles
- confiabilidad
  - se puede considerar que los datos representan información verídica

Atributos de calidad

- consistencia
  - no hay contradicciones entre distintos datos almacenados
- corrección
  - los datos representan la situación real
- seguridad/privacidad
  - los datos cumplen con los requerimientos de privacidad adecuados de acuerdo a la reglamentación nacional-internacional / criterios éticos
  - los datos son sólo accesibles por los usuarios autorizados

Calidad de datos

- No existen datos perfectos
- Es necesario priorizar las calidades deseadas

Temas

- **Calidad de datos**
  - Introducción
  - Definición
  - (Ejercicio e Ingeniería reversa)
  - Problemas habituales
  - Causas
  - Atributos de calidad
  - **Diagnóstico**
  - Corrección y prevención
- Algunas Definiciones
- Ley de protección de datos personales
- Open Data (Datos abiertos)

¿Cuán buena es la calidad de los datos?

¿Cómo determinar cuán buena es la calidad de los datos? (5 min.)

¿Cómo determinar cuán buena es la calidad de datos?

- Entender cuáles son los **datos críticos**
- Para estos determinar las dimensiones de interés

Para esto es necesario: **relevar**

Luego: **construir y ejecutar métricas** que permiten cuantificar la calidad de los datos

Relevamiento

Objetivos:

- determinar datos críticos, ciclo de vida del dato
- determinar atributos de interés

Tareas:

- identificar stakeholders: CCC (creator, consumer, custodian)
- leer documentación sobre los sistemas, sobre el negocio, y estudiar modelos de datos
- hacer cuestionarios tendientes a determinar cuáles son los datos críticos, cuál es el ciclo de vida del dato, cuáles son los atributos de interés y los problemas habituales.
- llevar adelante los cuestionarios con cada uno de los stakeholders identificados.

Necesidad de compromiso por parte del cliente - la organización

Elaboración de métricas

Para los datos críticos y las dimensiones de interés, armamos métricas para cuantificar cuán grave es el problema.

Una posible técnica: **GQM. Goal Question Metric** (Objetivo, Pregunta, Métrica)

Goal: definimos un objetivo
Pregunta: planteamos una o más preguntas, cuya respuesta nos permitirá saber si se satisface el objetivo
Métrica: planteamos una o más métricas -para cada una de las preguntas-, cuya ejecución nos permitirá responder las mismas

Bobrowski, Mónica, Martina Marré, and Daniel Yankelevich. **"A NEAT Approach for Data Quality Assessment: Industrial Experiences using a DQ Assessment Methodology."** In *Information and Database Quality*, pp. 135-162. Boston, MA: Springer US, 2002.

Elaboración de métricas

GQM. Goal Question Metric

Por ej. Departamento. Completitud.
G: El nombre del departamento debería estar completo
Q: Cuál es la proporción de registros que tienen el campo departamento vacío?
M1: Cantidad de registros con campo departamento vacío/cantidad total de registros
M2: proporción de id de departamento que tienen su nombre departamento vacío

Podemos ejecutar las métricas con SQL.

Los objetivos también podrían establecerse sobre **el modelo de datos**. Las respuestas podrían necesitar una revisión del modelo y ser sí o no.

Ejercicio - Consigna

- ✓ Conformar grupos de 3 integrante
- ✓ Usar la base de datos de Dengue de 2020 correspondiente al Registro del Sistema Nacional de Vigilancia de la Salud 2.0 (campus)-> Vigilancia de Dengue y Zika - 2020 (.xls)
- ✓ Aplicar la técnica GQM para evaluar la calidad de datos de dicha fuente

Diagnóstico

A partir de los resultados de las ejecuciones de las métricas y del relevamiento podemos determinar problemas existentes y sus causas.

Temas

- **Calidad de datos**
  - Introducción
  - Definición
  - (Ejercicio e Ingeniería reversa)
  - Problemas habituales
  - Causas
  - Atributos de calidad
  - Diagnóstico
  - **Corrección y prevención**
- Algunas Definiciones
- Ley de protección de datos personales
- Open Data / Datos abiertos

Resultados

A partir del diagnóstico: Conclusiones y propuestas de mejora.
Se trata no sólo de corregir, sino principalmente de prevenir
Posibles correcciones en:

- instancia
- modelo de datos
- procesos
- capacitación
- software

Herramientas para detección de problemas de calidad de datos

Existen muchas herramientas para automatizar detección de:

- textos parecidos (soundex, keyboard distance, edit distance, ..., uso de diccionarios).
- datos nulos
- problemas de integridad referencial

Temas

- Calidad de datos
- **Algunas Definiciones**
- Ley de protección de datos personales
- Open Data / Datos abiertos

¿Quién cuida los datos?

¿Los datos son valiosos, quién cuida su uso, su acceso, cómo?

¿Hay legislación al respecto?

5 minutos

Algunas definiciones

- Necesidad de definir términos
- Data management - Gestión de datos
- Data Governance - Gobernanza de datos
  - Data Owners-Dueño de los datos
  - Data Steward-Administrador de datos
- Algunas regulaciones sobre Datos

Algunas definiciones

- Necesidad de definir términos.
  - Ej. Datos vs. información
- Data management - Gestión de datos
  - La gestión de datos es el proceso de almacenamiento, organización y mantenimiento de los datos creados y recopilados por una organización.

Algunas definiciones

- Gobernanza de datos - Data Governance.
  - Concepto de gestión de datos relacionado con la capacidad que permite a una organización garantizar que exista una alta calidad de datos durante todo el ciclo de vida de los datos y que se implementen controles de datos que respalden los objetivos comerciales. Incluye el establecimiento de procesos para garantizar una gestión de datos eficaz en toda la empresa, la responsabilidad por los efectos adversos de la mala calidad de los datos y la garantía de que los datos que tiene una empresa puedan ser utilizados por toda la organización.
  - Las iniciativas de gobernanza de datos pueden estar impulsadas por el deseo interno de mejorar la calidad de los datos o por **regulaciones externas**.
  - Abarca a las personas, los procesos y tecnologías de la información necesarios para crear un manejo adecuado de los datos de una organización.

  Adaptado de Wikipedia

Algunas definiciones

- Data Owner- Dueño de los datos
- Data Steward- Administrador de datos
  - Un administrador de datos es un rol que garantiza que se sigan los procesos de gobierno de datos y que se cumplan las pautas, además de recomendar mejoras a los procesos de gobierno de datos.

Algunas definiciones

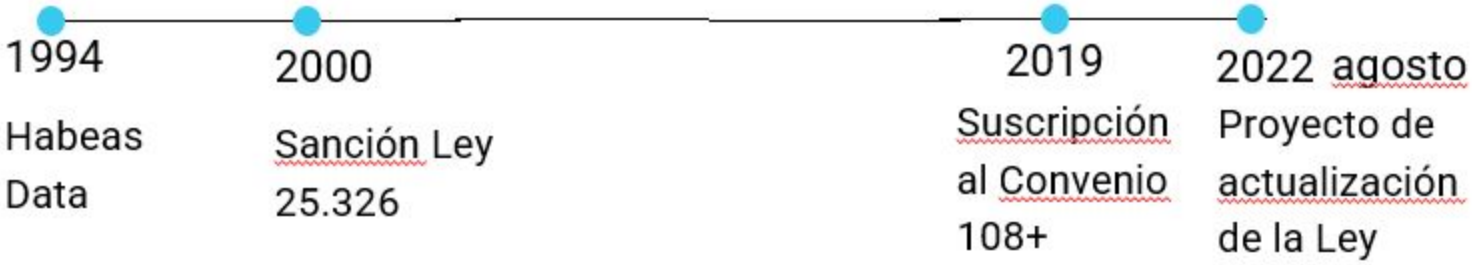
- Algunas Regulaciones
  - **Sarbanes-Oxley Act (SOX)**
    - La Ley Sarbanes-Oxley de 2002 es una ley federal de los Estados Unidos que exige **ciertas prácticas en el mantenimiento de registros financieros y la presentación de informes para las empresas**.
  - **HIPAA**
    - La Ley de Portabilidad y Responsabilidad del Seguro Médico de 1996 (HIPAA -Health Insurance Portability and Accountability Act–) es una ley federal que requirió la creación de **estándares nacionales para proteger la información confidencial de salud del paciente** para que no se divulgue sin el consentimiento o el conocimiento del paciente.
  - **GDPR**
    - La Legislación General de Protección de Datos es una ley de la Unión Europea de 2018, que **rige la forma en que se pueden usar, procesar y almacenar datos personales** (información sobre una persona viva e identificable).

Temas

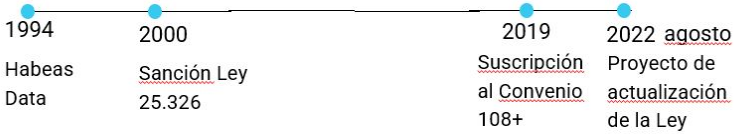
- Calidad de datos
- Algunas Definiciones
- **Ley de protección de datos personales**
- Open Data (Datos abiertos)

Protección de datos personales

Proceso normativo vinculado a la Protección de Datos Personales en Argentina



Protección de datos personales



Habeas Data (Constitución Nacional). Implica el derecho que tiene toda persona de saber cuáles y cómo se encuentran registrados sus datos en toda clase de registros.

Ley de Protección de datos personales 25.326. Establece entre otros qué son los datos sensibles, la finalidad de los registros. Obliga el registro de BD personales.

Convenios 108 y 108+. Convenio internacional al que adhirió Argentina. Relacionado con obligaciones de los Estados respecto al tratamiento automatizado de datos personales.

Otros instrumentos internacionales de PDP

Otros instrumentos internacionales:

- **RGDP**. Estándar internacional más garantista en la materia
- HIPPA
- Estándares de Protección de Datos Personales para los Estados Iberoamericanos-2017
- Recomendación sobre la Ética de la Inteligencia Artificial -2021
- Convenio 108 y su actualización (Convenio para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal. Europa)
- Reglamento Europeo de Protección de Datos personales
- Ley General de Protección de Datos Personales de Brasil-2018
- Ley Orgánica de Protección de Datos Personales de Ecuador- 2021
- Proyecto de reforma de la ley 25. 326 de Argentina - 2018
- Modificación a la Ley 19.628 de Protección de Datos Personales en Chile
- Proyecto de ley de Protección de Datos Personales en Paraguay

Nuevo proyecto de protección de datos personales

Charla: Nuevo proyecto de reforma de la ley de protección de datos personales: miradas y lineamientos para su implementación
Anastasia Dozo. https://www.youtube.com/watch?v=-seB4LVZ8Gs&t=8542s.
Duración: 36 minutos.

Protección por diseño: Desde la etapa de diseño de los sistemas, se tiene que tener en consideración la protección de datos personales.

Temas

- Calidad de datos
- Algunas Definiciones
- Ley de protección de datos personales
- **Open Data (Datos abiertos)**

¿Cómo consigo los datos?

¿Qué saben sobre datos abiertos?

5 minutos

Calidad de datos - Disponibilidad - Open data

Datos digitales disponibles de forma abierta. Que cualquier persona desde cualquier lugar puede:

-acceder

-utilizar con cualquier propósito y

-compartir

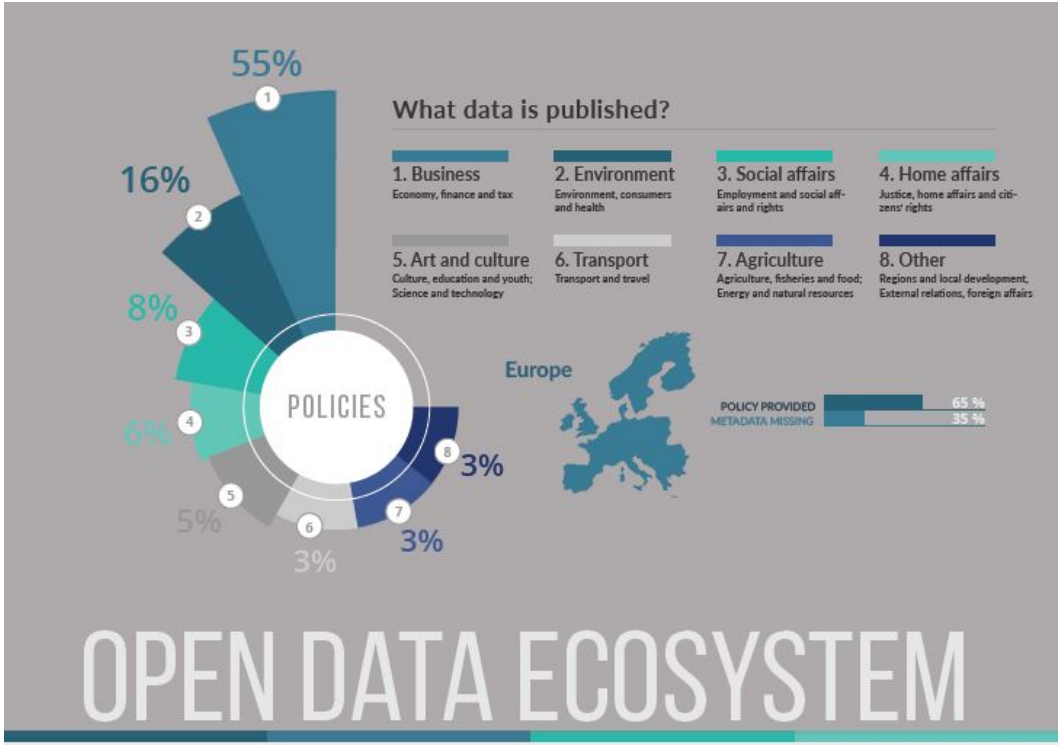
Open data

Tienen que satisfacer los siguientes requisitos:

- estar disponibles en la web
- tener un formato comprensible
- tener licencias, que permitan el uso libre de la información, incluyendo fines comerciales

Open data

A summary visualisation of the Open Data Ecosystem
(Source: OpenDataMonitor, 2016). Open Data Incubator



https://opendataincubator.eu/odine-startups-on-the-future-of-open-data/

Ley N° 27.275 de Acceso a la Información Pública (Septiembre, 2016)

Tiene por objeto garantizar el efectivo ejercicio del derecho de acceso a la información pública, promover la participación ciudadana y la transparencia de la gestión pública, y se funda en los siguientes principios:

. Presunción de publicidad
. Transparencia y máxima divulgación
. Informalismo
. Máximo acceso
. Apertura
. Disociación
. No discriminación.
. Máxima premura
. Gratuidad
. Control
. Responsabilidad
. Alcance limitado de las excepciones
. In dubio pro petitor
. Facilitación
. Buena fe

Gobierno abierto

Iniciativa que se fundamenta en la transparencia, la rendición de cuentas, la participación ciudadana, las tecnologías, el acceso a la información pública, los datos abiertos, la colaboración, la innovación, la eficiencia, la calidad de los servicios públicos, entre otros.

RISP Reutilización de Información en el Sector Público

Primer Plan de Acción Nacional de Gobierno Abierto (2013-2015)
Segundo Plan Nacional de Gobierno Abierto (2015-2017)
Tercer Plan de Acción Nacional de Gobierno Abierto (2017-2019). Link.

Principios para que los datos publicados puedan considerarse RISP según la CTIC y W3C

1. Datos completos.
2. Datos primarios.
3. Datos accesibles.
4. Datos proporcionados a tiempo.
5. Datos procesables.
6. Datos no discriminatorios.
7. Formatos no propietarios.
8. Datos libres de licencias.

Resumen

- DQ, ROI (retorno de la inversión), A quién le genera inconvenientes, Posibles causas, Diagnósticos, Estrategias de mejora
- Ley de protección de datos personales
- Datos abiertos

Tareas

- Hacer guía de Calidad de Datos.
- Leer Habeas Data.
- Leer la ley Argentina sobre protección de datos personales. Link
- Leer convenio 108. Link.
- Leer sobre GDPR. Link 1, Link 2.
- Ver video de JAIIO: Nuevo proyecto de reforma de la ley de protección de datos personales: miradas y lineamientos para su implementación. Link.

Referencias

- English, ‘Improving Data Warehouse and Business Information Quality‘, John Wiley & Sons, 1999
- Piatini, Calero, Genero (eds.): ‘Information and Database Quality, Kluwer‘, 2001. Cap 7: Bobrowski, Marré, Yankelevich, A NEAT Approach for Data Quality Assessment
- Redman (1996), ‘Data Quality for the Information Age’, Artech House.
- Wang, Strong, and Guarascio, ‘Beyond Accuracy: What data quality means to data consumers’, Total Data Quality Management Program, 1996.
- Las incluídas para lectura en la Tarea.