Calidad de datos

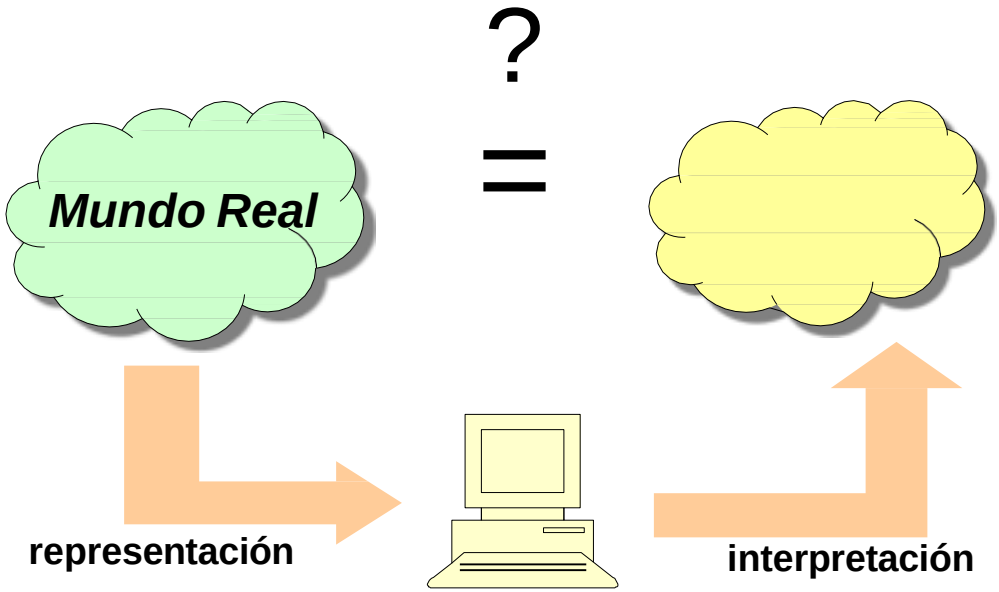
Bases de datos para Ingeniería, 2024

Lorena Etcheverry (lorenae@fing.edu.uy)
Instituto de Computación, FING, UdelaR

¿porqué nos preocupa la calidad de los datos?

Sobre los datos

Los datos representan objetos del mundo real en un formato adecuado para ser almacenado, recuperado y elaborado por un procedimiento de software y comunicado a través de una red.



Artificial Intelligence and Bad Data https://towardsdatascience.com/artificial-intelligence-and-bad-data-fbf2564c541a

Data in the real world is messy and dirty

Incompletos: falta de valores en atributos, falta de atributos de interés, o que sólo contienen datos agregados

- ejemplo: ocupación=“”

Con ruido: contienen errores (ortográficos, fonéticos o de digitación, transposición de palabras, valores múltiples en un campo de formulario) o contienen *outliers*

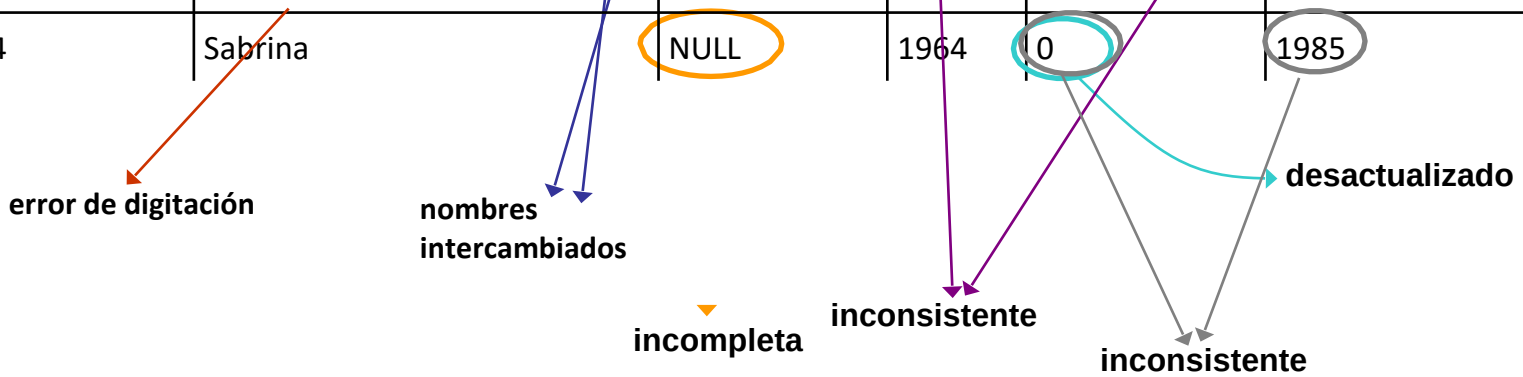
- ejemplo: sueldo=“-10”

Inconsistentes: contienen discrepancias entre los valores, o en el uso de códigos o nombres (sinónimos y apodos, variaciones en prefijos y sufijos, abreviaturas, truncados, iniciales)

- ejemplo: Edad=“42” y Fecha de nacimiento =“03/07/1997”
- ejemplo: algunas direcciones usan Av. y otras Avda. para abreviar avenida.
- discrepancia entre registros duplicados casi iguales

Ejemplos de problemas de calidad de datos

Código	Título	Director	Año	Cant-remakes	Ultimo-año-remake
1	Casablanca	Weir	1942	3	1940
2	La sociedad de los poetas Muertos	Curtiz	1989	0	NULL
3	Vacaciones en Rma	Wylder	1953	0	NULL
4	Sabrina	NULL	1964	0	1985



¿ Dónde se generan los problemas de calidad en los datos?

- Se generan a lo largo del ciclo de vida de los datos:
  - Producción de los datos
  - Procesamiento
  - Almacenamiento
  - Utilización

Algunas causas de la mala calidad

Producción de los datos

- Recolección de datos mediante ingreso humano
- Problemas sistemáticos con la recolección de datos
- Diferentes fuentes con representaciones diferentes del mismo objeto de la realidad
- No mantenimiento al día de los datos
- Ausencia de un responsable de los datos y de su calidad

Procesamiento

- Transformaciones a otras estructuras y formatos
- Cálculos con datos de entrada, como resúmenes y cálculos de indicadores
- Unión de datos provenientes de varias fuentes

Algunas causas de la mala calidad (cont)

Almacenamiento

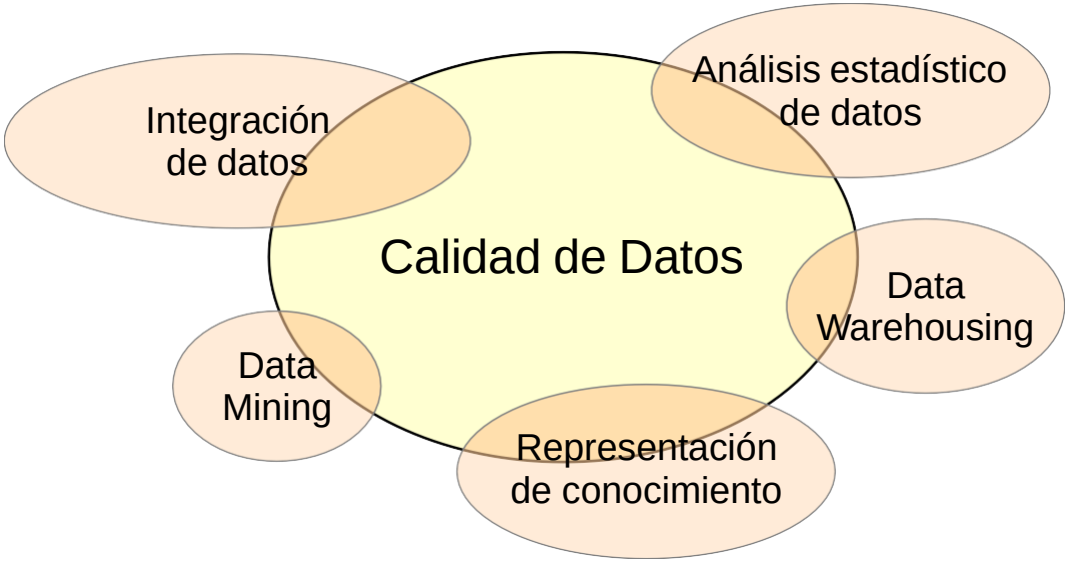
- Formatos diferentes
- Ausencia de formatos definidos
- Bases de datos mal diseñadas

Utilización

- Capacidad de análisis y procesamiento insuficiente
- Cambios en los requerimientos de calidad
- Uso equivocado de los datos, por mala interpretación o aplicación fuera de contexto
- Problemas de seguridad y acceso
- Mal diseño de los sistemas que procesan los datos para su análisis posterior

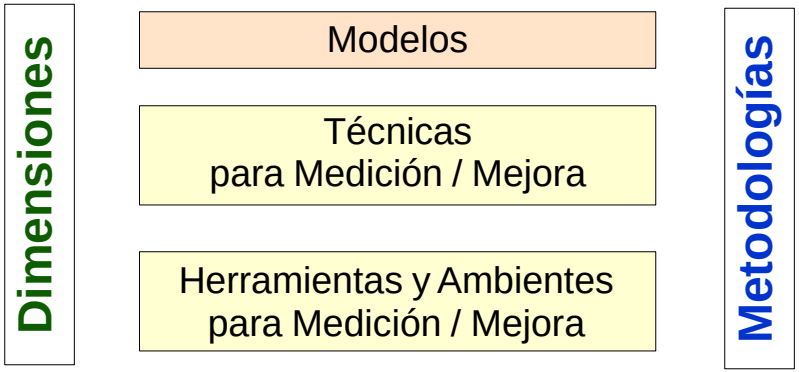
Líneas de Investigación

- Temas / Áreas de Investigación relacionados con Calidad de Datos



Líneas de Investigación

- Data Quality Management System (Sistema de Gestión de la Calidad)
  - Conjunto de técnicas, servicios y herramientas para manejar la calidad de los datos en una organización o varias cooperando.



Roles en Calidad de Datos

- **Chief Data Officer (CDO)** [wikipedia]
  - Es el responsable de la gobernanza y utilización de la información de toda la empresa, como un activo, a través del procesamiento, análisis, minería y comercio de datos.
  - Reporta al CEO
  - Es un gerente ejecutivo

- 90% de las grandes organizaciones tendrá un CDO en 2019, según Gartner
  - La carrera por competitividad y eficiencia a través de utilizar la información como un activo está llevando a un crecimiento abrupto de la cantidad de CDOs.

- A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data [Lee 2014]

P6

INFORMATION QUALITY MANAGEMENT MATURITY GRID

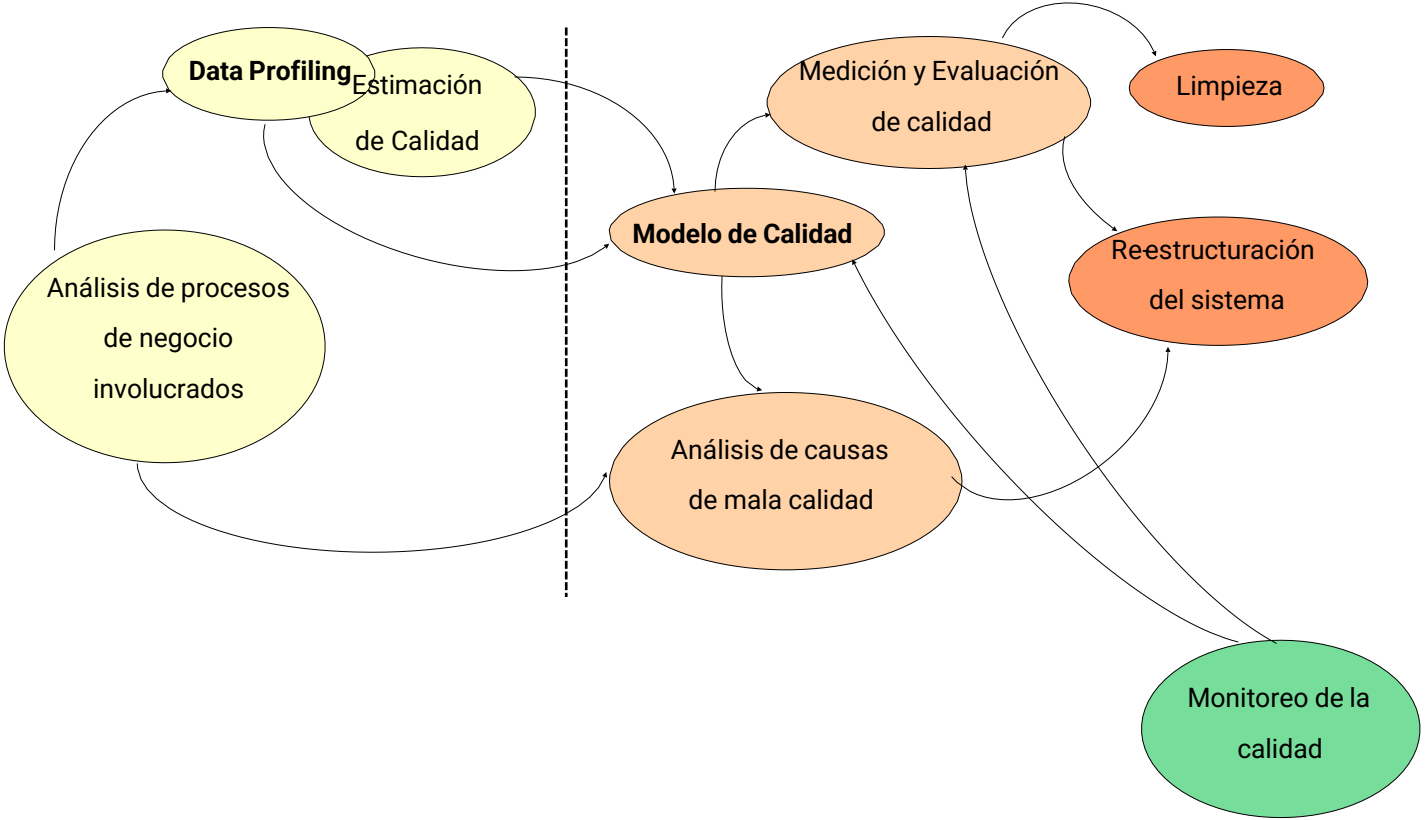
Measurement Categories	Stage 1: Uncertainty (Ad hoc)	Stage 2: Awakening (Repeatable)	Stage 3: Enlightenment (Defined)	Stage 4: Wisdom (Managed)	Stage 5: Certainty (Optimizing)
1. Management understanding and attitude	No comprehension of information quality as a management tool. Tend to blame Data Management or I/S org for "information quality problems" or vice versa.	Recognizing that information quality management may be of value but not willing to provide money or time to make it all happen.	While going through information quality improvement program learn more about quality management; becoming supportive and helpful.	Participating. Understand absolutes of information quality management. Recognize their personal role in continuing emphasis.	Consider information quality management an essential part of company system.
2. Information quality organization status	"Data" quality is hidden in application development departments. Data audits probably not part of organization. Emphasis on correcting bad data.	A stronger information quality role is "appointed" but main emphasis is still on correcting bad data.	Information quality organization exists, all assessment is incorporated and manager has role in development of applications.	Information quality manager reports to CIO; effective status reporting and preventive action. Involved with business areas.	Information quality manager is part of management team. Prevention is main focus. Information quality is a thought leader.
3. Information quality problem handling	Problems are fought as they occur; no resolution; inadequate definition; lots of yelling and accusations.	Teams are set up to attack major problems. Long-range solutions are not solicited.	Corrective action communication established. Problems are faced openly and resolved in orderly way.	Problems are identified early in their development. All functions are open to suggestion & improvement.	Except in the most unusual cases, information quality problems are prevented.
4. Cost of information quality as % of revenue	Reported: unknown Actual: 20%	Reported: 5% Actual: 18%	Reported: 10% Actual: 15%	Reported: 8% Actual: 10%	Reported: 5% Actual: 5%
5. Information quality improvement actions	No organized activities. No understanding of such activities.	Trying obvious "motivational" short-range efforts.	Implementation of the 14 point program with thorough understanding and establishment of each step.	Continuing the 14 point program and starting to optimize.	Information quality improvement is a normal and continued activity.
Summation of company information quality posture	"We don't know why we have problems with information quality."	"Is it absolutely necessary to always have problems with information quality?"	"Through management commitment and information quality improvement we are identifying and resolving our problems."	"Information quality problem prevention is a routine part of our operation."	"We know why we do not have problems with information quality."

Adapted from P. B. Crosby
Quality Management Maturity Model

IQMM® is a registered trademark of Information Impact Int'l L. English, *Improving Data Warehouse and Business Information Quality*, pg. 428

 © INFORMATION IMPACT Confidential & Proprietary

Gestión de la calidad en SI



Data Profiling

Data Profiling o Exploratory Data Analysis

- Primera aproximación al conocimiento sobre los datos del SI / dataset que queremos evaluar
  - Su estructura (si los metadatos son consistentes con los datos)
  - Sus relaciones
  - Su volumen
  - Sus problemas y frecuencia de los mismos
  - Patrones que se cumplen
- Algunas técnicas, como
  - Estadísticas básicas
  - Análisis de metadatos
  - Análisis de patrones
  - Detección automática de foreign keys

Data Profiling

- Análisis de atributos solapados de diferentes relaciones
  - Redundancias, claves foráneas
- Valores faltantes o erróneos
  - Cardinalidad actual vs. cardinalidad esperada (cant. clientes)
  - Frecuencia de valores nulos, maximo/minimo, etc.
- Duplicados
  - Número de tuplas vs. cardinalidad del dominio del atributo
- Claves difusas y dependencias funcionales difusas
  - Restricciones de integridad que no están explícitamente definidas pero que son satisfechas en la mayoría de los casos (un atributo que es clave, dependencias funcionales)

Ejemplos de Data Profiling con SQL

```
SELECT MIN(edad), MAX(edad), COUNT(DISTINCT edad)
FROM Empleados;

SELECT ciudad, COUNT(*) AS cant
FROM Clientes
GROUP BY ciudad ORDER BY cant;

SELECT COUNT(DISTINCT C1.ciudad)
FROM Clientes C1, Clientes C2
WHERE C1.ciudad = C2.ciudad AND
        C1.pais <> C2.pais;
```

Más ejemplos de Data Profiling con SQL

Estudiantes (ci-est, nombre, email, telefono, direccion, fnac)

Creo que el email no se repite casi nunca, intento verificar esto

```
SELECT COUNT(DISTINCT E1.email)
FROM Estudiantes E1
WHERE E1.email in
    (SELECT E2.email
    FROM Estudiante E2
    GROUP BY E2.email
    HAVING COUNT(*) > 1)
```

Y más ejemplos de Data Profiling con SQL

Actividades (ci-est, tipo-act, fecha, carrera, asignatura, instituto)

¿ Se cumple la dependencia funcional asignatura, carrera → instituto ?

```
SELECT DISTINCT A1.asignatura, A1.carrera
FROM Actividades A1, Actividades A2
WHERE A1.asignatura = A2.asignatura and
                A1.carrera = A2.carrera and
                A1.instituto <> A2.instituto
```

Ataccama	DQ Analyzer, Data Quality Center, DQ Issue Tracker, DQ Dashboard
Datactics	Data Quality Platform, Data Quality Manager, Master Record Manager
DataMentors	DataFuse, ValiData, NetEffect
Human Inference	HIquality Suite, HIquality Name Worldwide, HIquality Identify, HIquality Data Improver, DataCleaner
IBM	InfoSphere Information Analyzer, InfoSphere QualityStage, InfoSphere Discovery
Informatica	Data Explorer, Data Quality, Identity Resolution, AddressDoctor
Information Builders/iWay	iWay Data Quality Center
Innovative Systems	i/Lytics Data Quality, i/Lytics Data Profiling, i/Lytics ProfilerPlus, FinScan
Oracle	Oracle Enterprise Dat Quality, Oracle Enterprise Data Quality for Product Data
Pitney Bowes Software	Spectrum Technology Platform
RedPoint (DataLever)	RedPoint Data Management
SAP	Data Quality Management, Information Steward, Data Services
SAS/DataFlux	Data Management Platform
Talend	Talend Open Studio for Data Quality, Talend Enterprise Data Quality
Trillium Software	Trillium Software System, TS Discovery, TS Insight, Trillium Software On-Demand
Uniserv	Data Quality (DQ) Explorer, DQ Batch Suite, DQ Real-Time Suite, DQ Real-Time Services, DQ Monitor
Melissa Data	Contact Zone
Datiris	Datiris Profiler
CloverETL	Address Doctor
Microsoft	Data Quality Services

Data profiling en Python y R

Data Profiling en Python

Jean-Nicholas Hould
April 13th, 2017

DATA MANIPULATION +3

Exploratory Data Analysis of Craft Beers: Data Profiling

In this tutorial, you'll learn about exploratory data analysis (EDA) in Python, and more specifically, data profiling with pandas.

EJERCICIO: completar este tutorial

https://www.datacamp.com/community/tutorials/python-data-profiling

Data Profiling en R

EJERCICIO:
o completar este otro tutorial

https://datascienceplus.com/blazing-fast-eda-in-r-with-dataexplorer/

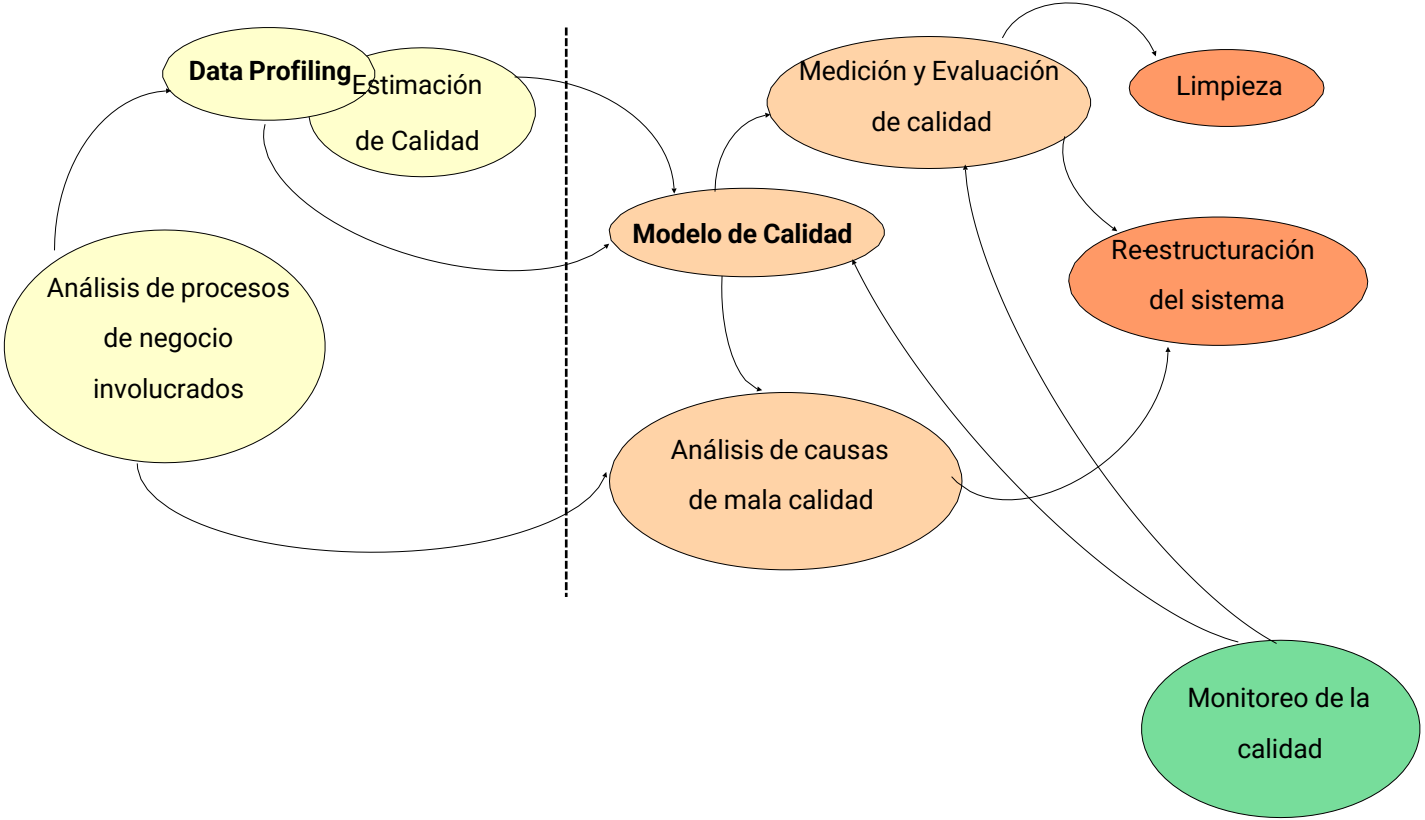
Bibliografía recomendada y referencias

- Ziawasch Abedjan, Lukasz Golab, Felix Naumann. **Data Profiling. SIGMOD 2017 Tutorial.** https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/2017/SIGMOD_2017_Tutorial_Data_Profiling.pdf

- Felix Naumann. **Data Profiling Revisited**. ACM SIGMOD Record 42.4 (2014): 40-49 https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/publications/2013/profiling_vision.pdf

Definiendo la calidad de los datos

Gestión de la calidad en SI



Modelo de Calidad de datos

Son artefactos que nos permiten definir que es la calidad de datos en cierto contexto

- Define **qué** características de calidad se manejan, sobre **qué** datos aplican, y **cómo** se miden esas características
- Para cada conjunto de datos se define un modelo de calidad particular
- Guía toda la gestión de la calidad de los datos

Pero, ¿ qué es la calidad?

Even though quality cannot be defined, you know what it is" – Robert Pirsig

Algunos sinónimos:

- Excelencia / valor
- Adecuación para su uso (*fitness for use*)
- Alcanzar o exceder las expectativas del consumidor

La Calidad de Datos y de Información es subjetiva,

depende del contexto, del uso, del consumidor, etc.

¿Qué es la Calidad de Datos?

Se la suele reducir a la exactitud de los datos (*accuracy*), sin embargo…

es un concepto multi-facético, donde existen diferentes **dimensiones.**

¿Qué pretendemos de los datos como consumidores?

- Que sean relevantes para su uso
- Que sean correctos y sin inconsistencias
- Que sean lo más actuales posible
- Que se vean en forma adecuada a sus aplicaciones
- Que sean fáciles de acceder

Problemas de calidad de datos

- ¿Qué problemas de calidad de datos de los SI que uds. manejan/mantienen encuentran en su trabajo cotidiano?
- ¿Qué problemas de calidad de datos han enfrentado en conjuntos de datos que uds usan?
- ¿Cómo clasificarían esos problemas según la lista nombrada anteriormente?

EJERCICIO: reflexionar sobre estas preguntas

La calidad como concepto multidimensional

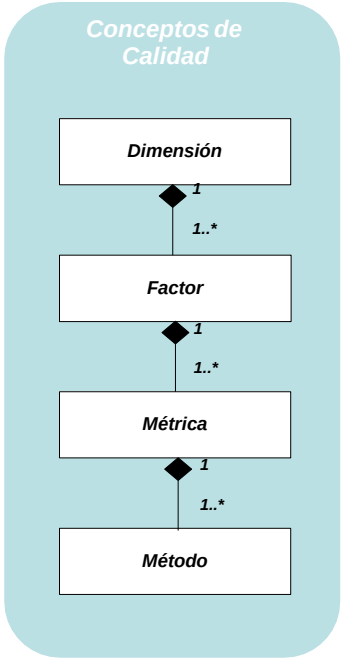
Multi-dimensionalidad de la calidad

- La calidad se caracteriza vía múltiples dimensiones o atributos que ayudan a calificar los datos.

- Conceptos de Calidad
  - Dimensión
  - Factor
  - Métrica
  - Método de Medición

Jerarquía de conceptos de calidad

- Las dimensiones representan las facetas de la calidad a alto nivel.
- Cada dimensión puede refinarse en un conjunto de factores que representan aspectos particulares.
- Cada factor puede medirse con varias métricas.
- Cada métrica puede implementarse con varios métodos de medición.



Multi-dimensionalidad de la calidad

- Dimensión de calidad:
  - Una dimensión captura una faceta (a alto nivel) de la calidad.
  - Ejemplos:
    - *Frescura*: los datos son recientes/actualizados.
    - *Exactitud*: los datos son exactos/correctos.
    - *Completitud*: disponemos de todos los datos.

- Factor de calidad:
  - Un factor representa un aspecto particular de una dimensión de calidad.
  - Ejemplo: Varios aspectos de la dimensión *Exactitud* son:
    - *Exactitud semántica*: si los datos representan entidades/estados del mundo real.
    - *Exactitud sintáctica*: si los datos no tienen errores sintácticos.
    - *Precisión*: si los datos tienen el suficiente nivel de detalle.

Medición de la calidad

- Métrica de calidad:
  - Define la forma de medir un factor de calidad
  - Se define con:
    - Un *nombre*
    - Una *descripción* (qué se mide)
      - Ej.: cantidad de valores nulos, cantidad de tuplas, tiempo transcurrido desde la última actualización
    - Las *unidades* de medición
      - Ej.: *tiempo* de respuesta en ms, volumen en GB, un valor entre 0 y 1, etc.
    - La *granularidad* de la medida
      - Fuertemente dependiente del modelo de datos
      - Modelo relacional: celda, tupla, columna, tabla, grupo de tablas, base de datos

Modelo Relacional

Tipo	Producto	Cant	PrecioUnit
Lacteos	Leche	5	1
Lacteos	Yogur	7	1.5
Bebidas	Agua Min	9	0.8

Medición de la calidad

- Método de medición:
  - Un método es un proceso que implementa una métrica.
  - Es el encargado de tomar una serie de medidas (correspondientes a una métrica) para una BD concreta.
  - Ejemplo: para medir el tiempo transcurrido desde la última actualización, se puede:
    - Usar timestamps de la BD
    - Acceder a los logs de actualización
    - Comparar versiones de la BD
    - ...
- Una misma métrica puede ser medida por diferentes métodos.

Ejemplo de conceptos de calidad

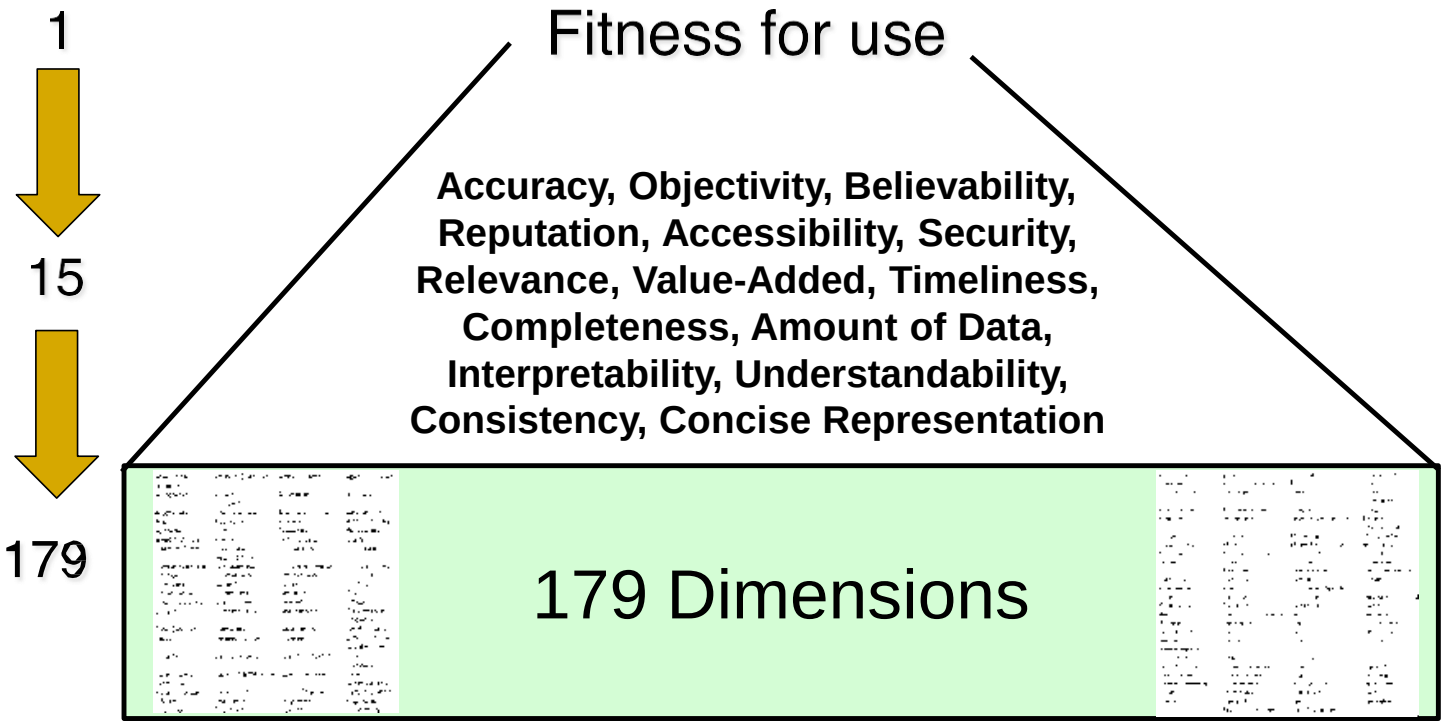
- Dimensión:
  - **Exactitud**: Concierne a cuan correcta y precisamente los datos del mundo real son representados en un sistema de información
- Factor:
  - **Exactitud sintáctica**: Indica qué tan libre de errores sintácticos están los datos
- Métricas:
  - **Exact. Sint. Booleana**: Un booleano que indica si un dato es sintácticamente correcto o no. *(Ej. un teléfono es correcto o no)*
  - **Desviación de exact. sint.**: La distancia a un dato considerado como sintácticamente válido (*Ej. Montevido, Mtdo*)
- Métodos:
  - **CheckRule**: Chequea si un dato satisface una regla de formato.
  - **CheckDictionary**: Chequea si un dato se encuentra en un diccionario.
  - **ComputeDistance**: Calcula la distancia entre un dato y el valor más cercano en un diccionario.

Agregación de medidas

- **Medida de calidad**
  - Es el valor obtenido al aplicar una métrica de calidad durante una medición (Ej., 0 o 1)

- **Granularidad**
  - En Modelo Relacional puede ser: Celda, Tupla, Atributo (columna), Tabla (o relación)
  - Análogamente en otros modelos
    - En XML/JSON podría ser: cada dato de determinados tags, o un documento.

- **Agregar**
  - Obtener a partir de la medida a determinado nivel de granularidad, la medida a un menor nivel de granularidad
  - Ej.: A partir de las medidas de las celdas, una medida para toda la tabla.


1
15
179
Fitness for use
Accuracy, Objectivity, Believability,
Reputation, Accessibility, Security,
Relevance, Value-Added, Timeliness,
Completeness, Amount of Data,
Interpretability, Understandability,
Consistency, Concise Representation
179 Dimensions

	WandWang 1996	WangStrong 1996	Redman 1996	Jarke 1999	Bovee 2001	Naumann 2002
Accuracy	S	S	S	S	S	S
Completeness	S	D	S	D%	S	S
Consistency	S		D%	S	S	
Representational Consistency		S	S			S
Timeliness	S	S		S	S	S
Currency	S		S	S	S	
Volatility	S			S	S	
Interpretability		S	D%	D%	S	S
Ease of Understanding/ Understandability		S				S
Reliability	D			D		
Credibility				D	D	
Believability		S				S
Reputation		S				S
Objectivity		S				S
Relevancy/ Relevance		S	S		D%	S
Accessibility		S		S	S	
Security/ Access Security		S		S		S
Value-added		S				S
Concise representation		S				S
Appropriate amount of data/amount of data		D	D			D
Availability				S		S
Portability			D	D		
Responsiveness/ Response Time				S		S

S-iguales
D-diferentes
D%-similares

Diferentes autores, diferentes criterios [Scannapieco 2002]

Algunas propuestas (1)

- ISO/IEC 25012, Dimensiones de calidad de datos para los Sistemas de Información

Inherentes	Inherentes y Dependientes	Dependientes
Exactitud Completitud Consistencia Credibilidad Actualidad	Accesibilidad Conformidad Confidencialidad Eficiencia Precisión Trazabilidad Entendibilidad	Disponibilidad Portabilidad Recuperabilidad

Algunas propuestas (2)

- C. Batini, M. Scannapieco. 2016. Data and Information Quality: Dimensions, Principles and Techniques.

Cluster	Tipo	Aspecto
Accuracy	Structural Accuracy	Syntactic accuracy
		Semantic accuracy
	Time-Related Accuracy	Currency
		Volatility
		Timeliness
Completeness	Completeness of Relational Data	Presence/absence and meaning of null values in an open/closed world assumption
	Completeness of Web Data	Completability
Accessibility	Accessibility	Accessibility
Consistency	Integrity Constraints	Intrarelational constraints
		Interrelational constraints
	Data Edits	Data editing

Vamos a ver en más detalle algunas dimensiones

- **Exactitud**
- Completitud
- Frescura
- Consistencia
- Unicidad

Exactitud (*accuracy*)

- Intuitivamente, la exactitud indica qué tan precisos, válidos y libres de errores están los datos:
  - ¿Estos datos son lo suficientemente precisos para nuestras necesidades?
  - ¿El nivel de detalle de los datos es adecuado?
  - ¿Estos datos se corresponden con el mundo real?
  - ¿Estos datos tienen errores? Y en tal caso, ¿los errores son tolerables?
  - ¿El formato de presentación de los datos es correcto? ¿Es estándar?

- La exactitud se relaciona con la corrección y la precisión con la que están representados los datos en un SI
  - Abarca aspectos de corrección que son intrínsecos de los datos y aspectos de representación (formato, precisión, etc.).

Factores de exactitud

- Exactitud semántica (semantic accuracy):
  - ¿Los datos de mi SI se corresponden con la realidad?
  - Interesa medir qué tan bien se representan los estados del mundo real en el SI.

  - Varios problemas de exactitud semántica:
    - Datos que no corresponden a ningún estado del mundo real (*mismembers*).
    - Datos que corresponden a un estado equivocado del mundo real.
    - Datos con errores en algunos atributos.
  - Ejemplo: Datos de un estudiante pueden referenciar
    - a una persona inexistente,
    - a una persona equivocada, o
    - a la persona correcta pero con algunos errores (ej. dirección)

Factores de exactitud

- Exactitud sintáctica (syntactic accuracy):
  - ¿Los datos de mi SI tienen errores sintácticos o de formato?
    - Valores mal escritos son difíciles de interpretar por un proceso
  - Interesa medir si los valores del SI corresponden a valores válidos del dominio (no importa si son los valores reales)

  - Varios problemas de exactitud sintáctica:
    - Errores de valores: Valores fuera de rango, errores ortográficos y de tipo.
      - Apellido: “Marínez” en lugar de “Martínez”
      - Edad: 338 años
    - Errores de estandarización: Valores que no tienen el formato esperado.
      - Sexo: “0” y “1” en lugar de “F” y “M”.
      - Precios: en moneda extranjera en lugar de pesos
      - Pesos: en gramos en lugar de kilos
    - Valores embebidos: Valores que corresponden a múltiples atributos
      - Dirección: embebe *calle número apto CP ciudad*.

Factores de exactitud

- Precisión (precision):
  - ¿Los datos de mi SI brindan el suficiente detalle?
  - Interesa medir qué tan detallados son los datos del SI.

  - Ejemplos:
    - Salario: “$10.000” vs. “$10.014” vs. “$10.013,88”
    - Fecha: “1977” vs. “julio de 1977” vs. “14/7/1977” vs. “14/7/1977 10:55:32.4”
    - Color: “Rojo” vs. “204R-51G-0B”
    - Cabello: “Castaño” vs. “Castaño claro cobrizo nº 5”
    - Dirección: “J.Herrera y Reissig 565, 11300, Montevideo” vs. “Montevideo”

Vamos a ver en más detalle algunas dimensiones

- Exactitud
- **Completitud**
- Frescura
- Consistencia
- Unicidad

Completitud (completeness)

- Intuitivamente, la completitud indica si el SI contiene toda la información de interés:
  - ¿El SI representa todos los objetos de nuestra realidad?
  - ¿Qué porción de la realidad está representada en el SI?
  - ¿Tenemos todos los datos que describen a nuestros objetos?
  - ¿Tenemos muchos valores nulos?

- La completitud recubre aspectos extensionales e intensionales del SI:
  - Extensional: La cantidad de entidades/estados de la realidad representados en el SI
  - Intensional: La cantidad de datos sobre cada entidad/estado del SI

Factores de completitud

- Cobertura (coverage):
  - ¿Cuántas entidades de la realidad contiene mi SI?
    - Mundo cerrado (close world): Una tabla contiene todos los estados de la realidad que ella describe.
    - Mundo abierto (open world): Una tabla puede contener sólo una parte de los estados de la realidad que ella describe.
  - Interesa medir la porción de los datos de la realidad contenidos en el SI.
  - Ejemplos:
    - De los clientes potenciales, ¿cuántos conozco?
    - ¿Qué porcentaje de las empresas están registradas en la DGI?

Factores de completitud

- Densidad (density):
  - ¿Cuánta info tengo sobre las entidades de mi SI?
  - Interesa medir cuánta info tengo y cuánta me falta sobre las entidades del SI.

  - Varias interpretaciones de la falta de valores (nulos):
    - Existen pero no los conozco (ej. No conozco el teléfono de Raquel).
    - Porque no existe (ej. Raquel no tiene teléfono).
    - No se si existe (ej. No se si Raquel tiene teléfono).

Vamos a ver en más detalle algunas dimensiones

- Exactitud
- Completitud
- Frescura
- **Consistencia**
- Unicidad

Consistencia (consistency)

- Intuitivamente, la consistencia captura la satisfacción de reglas semánticas definidas sobre los datos:
  - ¿Los datos satisfacen las reglas de dominio?
  - ¿Las dependencias funcionales y referenciales se satisfacen?
  - ¿Hay contradicciones entre los datos?

- Pueden ser reglas de integridad para una BD o reglas de los usuarios
  - Reglas de integridad: son propiedades que deben satisfacer todas las instancias de una BD.
  - Reglas de usuarios: no implementadas en la BD pero necesarias para una aplicación.

Factores de consistencia

- Integridad de dominio
  - Satisfacción de reglas sobre el contenido de un atributo.
    - Ej. edad entre 0 y 120 años.

- Integridad intra-relación
  - Satisfacción de reglas entre atributos de una misma tabla.
  - Reglas más típicas:
    - Dependencias de clave y de unicidad
    - Dependencias funcionales
    - Dependencias de valores. Ej. Edad = Year (now() – FechaNacimiento)
    - Expresiones condicionales (edits). Ej. EstadoCivil = “casado” Edad ≥ 14

- Integridad inter-relación
  - Satisfacción de reglas entre atributos de varias tablas.
  - Reglas más típicas:
    - Dependencias de inclusión (clave foránea, integridad referencial)

- Interesa medir qué tan bien se satisfacen las reglas de integridad

Problemas de calidad de datos

- ¿Qué problemas de calidad de datos de los SI que uds. manejan/ mantienen encuentran en su trabajo cotidiano?

- ¿Qué problemas de calidad de datos han enfrentado en conjuntos de datos que uds usan?

EJERCICIO: Identificar las dimensiones y los factores de calidad donde clasificaría cada uno de los problemas antes descriptos

Dimensiones tradicionales - BigData

	Datos Transaccionales	Big Data o Social Media Data
ACCURACY	Utilizamos un referencial para medirla.	En general no es posible tener un referencial. Crowd-sourcing?
CONSISTENCY	Controles en el ingreso de datos. Restricciones de integridad y triggers para reglas de negocio (BDs relacionales)	Imposible controlar ingreso de datos. BDs NoSQL no tienen funcionalidades para eso, y muy costoso implementarlo
TIMELINESS	Tiempo transcurrido entre generación del dato y uso del mismo	Típicamente describe eventos en tiempo real. Esta información es muy sensible al momento en que se genera / lee.
COMPLETENESS	De gran relevancia. Cobertura y Densidad.	No tenemos esquema para medir densidad. Cobertura depende del contexto de uso, para qué y con respecto a qué.

Bibliografía

- Carlo Batini, Monica Scannapieco. Data and Information Quality. Springer. ISBN: 978-3-319-24104-3. 2016.
- Thomas C. Redman. Data Quality for the Information Age. 1996 Artech House Inc., ISBN 0-89006-883-6
- G. Shankaranarayanan y R. Blake. From Content to Context: The Evolution and Growth of Data Quality Research. J. Data and Information Quality, vol. 8, n.º 2, p. 9:1–9:28, 2017.
- Jack E. Olson. Data Quality. The Accuracy Dimension. Morgan Kaufmann Publishers, Elsevier. 2003. ISBN-10 1-55860-891-5
- W. Eckerson. Data Warehouse Institute Survey on Data Quality. Proceedings of the Seventh International Conference on Information Quality (ICIQ-02).
- Larry English. The TIQM® Quality System for Total Information Quality Management: Business Excellence through Information Excellence. MIT Information Quality Industry Symposium, 2009.
- Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, “AIMQ: a methodology for information quality assessment,” Information & management, vol. 40, no. 2, pp. 133–146, 2002.

Bibliografía (cont)

- S. E. Madnick, R. Y. Wang, Y. W. Lee, and H. Zhu, “Overview and Framework for Data and Information Quality Research,” J. Data and Information Quality, vol. 1, no. 1, pp. 2:1–2:22, Jun. 2009.
- Felix Naumann, Kai-Uwe Sattler. Information Quality: Fundamentals, Techniques and Use. EDBT Tutorial, Munich, 2006.
- D. M. Strong, Y. W. Lee, and R. Y. Wang, “Data quality in context,” Commun. ACM, vol. 40, no. 5, pp. 103–110, May 1997.
- R. Y. Wang and D. M. Strong, “Beyond accuracy: What data quality means to data consumers,” Journal of management information systems, pp. 5–33, 1996.
- M. Scannapieco and T. Catarci, “Data quality under a computer science perspective,” Archivi & Computer, vol. 2, pp. 1–15, 2002.
- B. Otto, K. M. Huner, and H. Osterle, “Identification of Business Oriented Data Quality Metrics,” presented at the ICIQ, 2009, pp. 122–134.
- Y. Lee, S. Madnick, R. Wang, F. Wang, H. Zhang. A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data. MIS Quarterly Executive, 2014.