Universidad Central “Marta Abreu” de Las Villas

Facultad Matemática Física y Computación

Departamento Ciencia de la Computación

Monografía

Calidad de datos

Colectivo de Autores

Edición: Claudia María Larrea Marín
Corrección: Marisleidy Pino Ortega

Editorial Feijóo, 2016

ISBN: 978-959-312-157-6

Editorial Samuel Feijóo, Universidad Central “Marta Abreu” de Las Villas, Carretera a Camajuaní, km 5 ½, Santa Clara, Villa Clara, Cuba. CP 54830

RESUMEN

La tendencia actual adquirida por la mayoría de las empresas modernas de almacenar cantidades crecientes de información ha provocado que la gestión de la calidad de datos se convierta en un proceso sumamente importante. En el presente trabajo se exponen las principales definiciones del término calidad de datos y su ciclo de vida a través de la caracterización de las cuatro fases que componen la metodología para la gestión total de la calidad de datos: definir, medir, analizar y mejorar. Además, se analiza la correspondencia que existe entre las dimensiones y las métricas de la calidad de datos, así como del perfilado y la limpieza de datos con estas fases.

Palabras claves: calidad de datos, dimensiones, limpieza de datos, metodología, métricas, perfilado de datos.

Tabla de contenido

INTRODUCCIÓN

Los datos son realidades concretas que se encuentran en el mundo, hechos atómicos que por sí solos carecen de significado (Batini and Scannapieco, 2006). Su calidad puede influir sustancialmente en tres vectores estratégicos de las organizaciones: interacción con el cliente, configuración de recursos y aprovechamiento del conocimiento (Venkatraman and Henderson, 1998). Por consiguiente, los datos son un recurso importante para la competitividad organizacional (Redman, 1998).

En las condiciones actuales en que los negocios son cada vez más dinámicos y competitivos, los datos adquieren un gran valor para mejorar su eficiencia. De ahí que el éxito del negocio y la satisfacción de los clientes dependen en gran medida de la calidad de los datos.

Para que un proceso de calidad de datos sea realmente eficaz, este debe ser repetible y fácil de entender, de manera que permita generar un ciclo de mejora y que cada vez que sea ejecutado se obtengan datos con mayor calidad. El proceso debe incluir el perfilado de los datos, la estandarización o normalización, la correspondencia y consolidación; pasos que en su mayoría generan reportes para dar seguimiento a los progresos y permitir la mejora continua de su calidad.

En la mayoría de las organizaciones y sistemas empresariales que comienzan a crear almacenes de datos integrados, los problemas de la calidad de datos aparecen ineludiblemente. Estudios de diversos grupos de investigadores en el mundo han coincidido en que las fallas de los proyectos de implementación de almacenes de datos se deben a un mantenimiento insuficiente de la calidad de datos, lo que conduce a errores en la toma de decisiones.

La baja calidad de datos afecta la reputación de las organizaciones al tomar decisiones basadas en datos erróneos y no fiables y propicia la acumulación de errores.

Por tal motivo garantizar y mantener una adecuada calidad en los datos es fundamental para cumplir metas y objetivos.

ASPECTOS TEÓRICOS SOBRE CALIDAD DE DATOS

El presente trabajo tiene como objetivo analizar los principales aspectos relacionados con el proceso de gestión de la calidad de datos, comenzando por las principales definiciones atribuidas al término desde el punto de vista de diferentes autores. Seguido de esto, se caracterizan las cuatro fases, metodología para la gestión total de la calidad de datos (TDQM, por sus siglas en inglés): definir, medir, analizar y mejorar. Finalmente, se analiza la correspondencia que existe entre las dimensiones y las métricas de la calidad de datos, así como del perfilado y la limpieza de datos con las cuatro fases mencionadas anteriormente.

Definición de calidad de datos

"Es difícil dar una definición universal de lo que significa calidad" (Bobrowski *et al.*, 1999). La noción de calidad se utiliza en el lenguaje cotidiano como un rasgo intangible, algo que puede ser subjetivamente juzgado, pero a menudo no es medido exactamente. Términos como buena o mala calidad son intensamente difusos y utilizados sin la intención de ser una ciencia exacta. Además la calidad posee carácter multidimensional e incluye un objeto de interés, el punto de vista de ese objeto y los atributos de calidad atribuidos al mismo. Esto puede hacer del término "calidad", un concepto confuso. Con el fin de discutir la calidad, poder evaluarla y mejorarla, esta tiene que ser definida (Wingkvist *et al.*, 2010).

Una de las definiciones más aceptadas para el término "calidad de datos" está dada por (Juran y Gryna, 1980), y retomada por diversos investigadores (Wang y Strong, 1996; Francalanci y Pernici, 2004; Chapman, 2005; Shankaranarayanan y Cai, 2006; Batini *et al.*, 2009; Sadiq *et al.*, 2011; Salomone *et al.*, 2011; Moges *et al.*, 2013; Yeganeh *et al.*, 2014), quienes coinciden en definirla como "datos adecuados para el uso" de los consumidores de datos. Esto significa que el usuario es el encargado de determinar si el nivel de calidad de un conjunto de datos usados en un contexto específico de acuerdo a determinados criterios de calidad satisface sus necesidades.

Un problema crucial es determinar cuáles son los criterios de calidad de datos que deben ser considerados para la selección de los datos en función del uso que se les da. Aún más, cuando se trata de fuentes de datos internas y externas, pues es necesario determinar un intercambio satisfactorio entre la cantidad y la calidad de datos recuperados (Abelló *et al.*, 2013). Por consiguiente, al evaluar la calidad de datos de las

fuentes encontradas, la atención debe centrarse en el análisis de cómo la calidad de datos actual afecta negativamente el análisis posterior de los datos y las decisiones que pueden tomarse a partir de los mismos.

La calidad de datos se ha estudiado ampliamente y desde diferentes puntos de vista durante las últimas décadas. El Instituto de Tecnología de Massachusetts (MIT, por sus siglas en inglés) ha sido uno de los pionero en investigaciones de calidad de datos desde que el programa de gestión total de calidad de datos (TDQM) fue iniciado en 1992 (Madnick y Wang, 1992; Madnick *et al*., 2009).

Ciclo de vida de la metodología TDQM

TDQM fue la primera metodología publicada en la literatura de calidad de datos (Wang, 1998). Esta metodología es el resultado de investigaciones académicas y ha sido exhaustivamente usada como guía para iniciativas de reingeniería de datos organizacionales (Batini *et al*., 2009). TDQM utiliza una versión adaptada del ciclo propuesto por W. E. Deming en el año 1986. Este ciclo se ha convertido en un tema clave en la literatura para la gestión total de la calidad (TQM, por sus siglas en inglés) (Wijnhoven *et al*., 2007) y se identifica por cuatro pasos principales: planificar, hacer, verificar y actuar (Deming, 1986).

Años más tarde, Wang (1998) realizó una analogía entre la TQM de productos físicos manufacturados y la TDQM, donde argumenta que la fabricación de un producto puede ser vista como un sistema de procesamiento que actúa sobre las materias primas para producir productos físicos. Análogamente, la fabricación de información puede ser vista como un sistema de procesamiento que actúa sobre los datos en bruto para producir productos de información (Koronios *et al*., 2005). Teniendo en cuenta esta analogía (ver Tabla 1), Wang propuso un sistema de fabricación de información, adaptando el método de Deming en las fases: definir, medir, analizar y mejorar los productos para aplicarlos luego al entorno de fabricación.

Tabla 1: Fabricación de productos contra fabricación de información

	Fabricación de productos	**Fabricación de información**
Entrada	Materiales sin pulir	Datos en bruto
Proceso	Línea de ensamblaje	Sistema de información
Salida	Productos físicos	Productos de información

Fuente: (Wang, 1998)

En la Figura 1 se ilustra el ciclo TDQM para una mejora continua y entrega de productos de información de alta calidad (Koronios *et al.*, 2005).

Figura 1: Ciclo de vida de la metodología TDQM

Basado en: (Wang, 1998)

Fase de definición

Según Wijnhoven *et al*., 2007) la fase de definición se compone de tres pasos, los cuales consisten en determinar las características de los productos de datos, los requisitos para los productos de datos y el proceso de fabricación de datos.

1. Determinar las características de los productos de datos

Para describir las características de un producto de datos, Wang (1998) utiliza dos niveles. En el nivel más alto se describen las características generales del producto de datos, y en el nivel más bajo se describe cada atributo del producto de forma individual. La Figura 2 muestra la diferencia entre las características de los productos de datos y los atributos de dichos productos (Wijnhoven *et al.*, 2007).

Figura 2: Datos del producto y datos de los atributos del producto

Fuente: (Wijnhoven *et al*., 2007)

No todos los productos de datos y los atributos pueden ser evaluados en un instante de tiempo. Wijnhoven *et al.* (2007) sugieren centrarse en los productos de datos y atributos cuyos problemas de calidad tienen mayor impacto.

2. Determinar los requisitos para los productos de datos

La metodología TDQM propone determinar cuáles de las dimensiones de calidad asociadas a los productos de datos (o atributos) (Wang y Strong, 1996) son las más importantes, atendiendo a preguntas sobre:

- El nivel de percepción de calidad en una dimensión, y
- El nivel esperado de calidad en una dimensión.

3. Determinar el proceso de fabricación de datos

El proceso de fabricación de los datos consiste en flujos de datos que van desde el proveedor hasta los datos del usuario, incluidas ciertas actividades de procesamiento y controles de calidad (Wijnhoven *et al.*, 2007). Según Wang (1998), el conocimiento del proceso de fabricación sirve como base para comprender mejor por qué ciertas dimensiones de calidad son importantes y cómo se dividen las responsabilidades sobre este proceso. Además, pueden detectarse en este paso los procesos y almacenamientos redundantes, los cuales a menudo conducen a errores e inconsistencias. Tener un proceso claramente definido ayuda en la gestión de calidad de datos. Una versión simplificada de un proceso de fabricación de datos es mostrado en la Figura 3 (Wijnhoven *et al.*, 2007).

Figura 3: Ejemplo de un sistema de fabricación de datos

Fuente: (Wijnhoven *et al.*, 2007)

Fase de medición

La fase de medición determina la calidad de las dimensiones identificadas en la fase de definición. Esta fase se basa en dos pasos fundamentales (Wijnhoven *et al.*, 2007).

1. Seleccionar las métricas adecuadas

Para evaluar una misma dimensión de calidad existen varias métricas, pero cuál escoger depende del contexto de análisis (caso de estudio). En la determinación de una métrica para una dimensión de calidad de datos, se debe tener en cuenta qué hay detrás de las reglas de negocio, normas ISO (acrónimo del inglés, *International Organization for Standardization*) y leyes que pueden haber contribuido a la importancia de las dimensiones identificadas (Wijnhoven *et al.*, 2007). Según Pipino *et al*., (2002), se identifican las siguientes tres formas de medición de dimensiones de calidad: relación simple, operación de mínimo (*min*) o máximo (*max*) y media ponderada. Usualmente es difícil determinar qué métrica utilizar en cada dimensión de calidad elegida en la fase anterior, por tanto, se sugiere usar como referente las pautas RUMBA propuestas por (Kovac *et al*., 1997), ver Tabla 2.

Tabla 2: Pautas RUMBA para las métricas

R	¿Es una métrica razonable (***R**easonable*)?
U	¿Es una métrica comprensible (***U**nderstandable*)?
M	¿Es una métrica medible (***M**easurable*)?
B	¿Es una métrica creíble (***B**elievable*)?
A	¿Es una métrica alcanzable (***A**chievable*)?

Fuente: (Kovac *et al*., 1997)

2. Medir y presentar los datos

La medición puede llevarse a cabo una vez que se determinan las métricas que se van a utilizar y los datos a medir. Por ejemplo, si existen 20 mil productos en la base de datos, se necesita un tamaño de muestra de aproximadamente 400 productos para alcanzar una fiabilidad de un 95 % con una posible variación de un 5 % (Moore y McCabe, 1989).

En la visualización de los resultados se pueden utilizar varios tipos de gráficos, por ejemplo, para identificar qué dimensiones son las causantes de que ocurran problemas en los resultados, se puede hacer uso de los diagramas de Pareto (Zahedi, 1995). En estos gráficos el eje horizontal muestra los errores en las diferentes dimensiones y el eje vertical muestra el porcentaje de los errores en comparación con el número total de errores. Esta es una manera fácil de mostrar que las dimensiones causan la mayoría de los problemas (Wijnhoven *et al.*, 2007).

Fase de análisis

El objetivo de esta fase es encontrar las causas fundamentales que conducen a los problemas de calidad en las diferentes dimensiones identificadas (Wang, 1998). Tres

métodos para encontrar dichas causas se han hecho populares en la actualidad: los diagramas de causa y efecto (CED, por sus siglas en inglés), los diagramas de interrelación (ID, por sus siglas en inglés), y los árboles de realidad actual (CRT, por sus siglas en inglés). Según (Doggett, 2005), el método CED es la herramienta más fácil para identificar las principales causas.

En esta etapa se deben responder las siguientes interrogantes: ¿cuál es el problema?, ¿cuándo pasó?, ¿por qué pasó? y ¿cómo afecta a los objetivos generales de la empresa? (Wijnhoven *et al.*, 2007).

Fase de mejora

Para la fase de mejora, (Wang, 1998) propone cuatro pasos, dentro de los cuales se encuentra el proceso de limpieza de datos. Actualmente existe un considerable conjunto de técnicas y algoritmos de limpieza de datos que se emplean en diversas fuentes con el propósito de mejorar su calidad sin afectar la naturaleza de los mismos. En ocasiones estas técnicas consumen bastante tiempo, presentan elevada complejidad computacional o no brindan soluciones exactas, por lo que esta área permanece activa dentro de la comunidad científica.

1. Generación de soluciones

Para este paso, (Wang (1998) propone el uso de la matriz de análisis de la fabricación de datos diseñada por (Ballou *et al.*, 1998) y presentada en la Figura 3, desde la cual se puede obtener la falta de controles de calidad e identificar posibles procesos y fuentes de datos redundantes.

2. Selección de soluciones

Para seleccionar las soluciones, es necesario conocer qué impacto pueden tener sobre la gestión de los datos. Se debe analizar el costo de diseñar e implementar una solución, los recursos requeridos y cómo reducir el riesgo de introducir ruido en los datos. Las prioridades para las soluciones pueden establecerse cuando la importancia de cada criterio de evaluación es conocido (Zahedi, 1995).

3. Desarrollo de un plan de acción

La solución seleccionada tendrá que ser transmitida en acciones concretas. En este paso se debe asegurar que todas las acciones sean ejecutadas. Un plan de acción del proyecto se puede configurar para realizar un seguimiento de todas las acciones y sus estados (Wijnhoven *et al.*, 2007).

4. Control de progresos

Con el paso control de progresos se puede conocer si ocurren deficiencias al llevar a cabo las acciones implementadas para el plan de acción desarrollado. En caso de fallo se comienza a ejecutar nuevamente el ciclo que encierra la metodología, o se espera un tiempo en caso que el progreso sea satisfactorio, o sea hasta que la mayoría de los datos estén limpios (Wijnhoven *et al.*, 2007).

El propósito de la metodología TDQM es entregar productos de información de alta calidad a los consumidores de datos. Su objetivo es facilitar la aplicación de la política general de la calidad de datos de una organización expresada formalmente por los altos directivos (Wang *et al.*, 1995).

Dimensiones de calidad de datos

En dependencia del contexto particular que se esté analizando usualmente interesa tener en cuenta únicamente los atributos (dimensiones) de calidad con cierta relevancia determinada por los usuarios, los especialistas del negocio o los consumidores de datos. Por ejemplo, la calidad de datos en el contexto de los sistemas de información se relaciona con los beneficios que pueden brindar a una organización. Por tanto, para obtener una medida precisa de calidad de datos se debe elegir qué dimensiones conviene considerar y qué cantidad de cada una de ellas se debe medir, de manera que se contribuya a la calidad como un todo (Bobrowski *et al.*, 1999).

Principales técnicas para identificar las dimensiones de calidad de datos

La identificación de dimensiones de calidad de datos en un contexto específico juega un papel fundamental en la fase de definición de la metodología TDQM.

Si no se identifican las dimensiones pertinentes, no se pueden abordar eficazmente los problemas de calidad de datos. Así, el primer objetivo de toda investigación o estudio práctico sobre este tema, es determinar las características de los datos que son importantes para los consumidores de datos, o que resultan convenientes para los mismos (Wang y Strong, 1996). Mientras que el término “adecuación al uso” capta la esencia de la calidad de datos, desde hace tiempo se ha reconocido que los datos son mejor descritos o analizados a través de múltiples atributos o dimensiones (Tayi y Ballou, 1998; Shankaranarayanan y Cai, 2006; Maydanchik, 2007). Sin embargo, a pesar de un amplio debate en la literatura de calidad de datos, no existe un único conjunto definido de dimensiones de calidad de datos que se pueda aplicar universalmente, porque dependen del contexto en que se estén analizando (Moges *et al.*, 2013).

Diferentes estudios han analizado las dimensiones de calidad de datos desde cierta perspectiva para una tarea específica. Por ejemplo, Zhu y Gauch (2000) evaluaron la calidad de datos de una página *web* en términos de un marco de calidad comprendido en las siguientes seis dimensiones: actualidad, disponibilidad, relación información-ruido, autoridad, popularidad, y cohesión; midiendo las dimensiones a través de las propiedades de las páginas *web*. Del mismo modo, Chen y Tseng (2011) evaluaron diferentes dimensiones de calidad de datos con el fin de calcular la calidad de la revisión de productos en línea realizada por los clientes; adoptando varias definiciones de diferentes dimensiones para el análisis de la calidad de los comentarios de productos en línea. Por ejemplo, definieron la objetividad como el grado en que un elemento de información es parcial, la cantidad adecuada de los datos como la medida en que el volumen de información en una revisión es suficiente para la toma de decisiones, y la integridad como la medida en que la información en una crítica es completa y abarca diversos aspectos de un producto. Además, identificaron la objetividad y la cantidad adecuada de la información como dimensiones de calidad de datos eficaces en la identificación de la calidad para la revisión de un producto e ineficaces para una completa evaluación a la hora de medir la calidad de una revisión por los clientes u otras partes (Moges *et al.*, 2013).

Por otro lado, hay una serie de estudios que identifican y definen las dimensiones de calidad de datos independientemente de la utilización de los datos, con el fin de facilitar la aplicabilidad general y la comparabilidad de sus dimensiones (Moges *et al.*, 2013). En este sentido, Wand y Wang (1996) basaron su definición de calidad de datos, auxiliados en la vista interna de los sistemas de información (producción de datos y los procesos de diseño del sistema), ya que este punto de vista es independiente del contexto. Este enfoque permite que una serie de definiciones de dimensiones de calidad de datos sean comparables entre aplicaciones. En primer lugar, identificaron varios criterios de un sistema del mundo real para posteriormente ser representado adecuadamente por un sistema de información. En base a estos criterios, definieron cuatro deficiencias nombradas: representación ambigua, representación incompleta, estados sin sentido y deficiencias de operación. Sobre la base de estas deficiencias, resumieron diferentes aspectos de calidad de datos en dimensiones completas, sin ambigüedades, con sentido y corregidas. Además, en el mismo estudio, se clasifican las diferentes dimensiones de calidad de datos de la literatura como vista interna (diseño u operación relacionada) o vista externa (uso o el valor relacionado), por medio del cual ambos puntos de vista se vuelven más refinados como cualquiera de los sistemas o datos relacionados con dimensiones de calidad. Dentro de la visión interna, la exactitud o precisión, el tiempo de vida o la actualidad, la fiabilidad, la integridad y la consistencia, son definidas como datos relacionados, mientras que la fiabilidad es

definida como una dimensión relacionada con el sistema. Por otra parte, en la vista externa, el tiempo de vida, la relevancia, el contenido, la importancia y la cantidad son definidas como datos relacionados, mientras que el tiempo de vida, la flexibilidad, el formato y la eficiencia son definidas como dimensiones relacionadas con el sistema (Moges *et al.*, 2013).

Del mismo modo, Wang y Strong, (1996) se analizan diversas dimensiones de calidad de datos desde varias perspectivas del usuario final, pero sin tener en cuenta el uso de los datos. Además, se utiliza como técnica para determinar y clasificar las dimensiones de calidad el diseño y aplicación de una encuesta a gran escala. Su análisis se inició mediante la recopilación de información de los usuarios, en relación con diversos descriptores de calidad de datos que resultaron en más de 100 ítems, los cuales fueron agrupados en la primera fase del estudio en 20 dimensiones y estas últimas fueron agrupadas en las siguientes cuatro categorías generales de calidad de datos: intrínseca (grado en el que los valores de los datos están en conformidad con los valores reales o verdaderos), contextual (medida en que los datos del usuario son aplicables a una tarea), representacional (grado en que los datos se presentan de una manera comprensible) y accesibilidad (grado en que los datos están disponibles o accesibles) Wang y Strong, (1996). En la segunda fase del estudio el total de dimensiones fue ajustado teniendo en cuenta los resultados de la primera fase, quedando finalmente 15 dimensiones clasificadas en las cuatros categorías mencionadas anteriormente, en el marco de trabajo propuesto por Wang y Strong (1996).

Este marco se reconoce por intentar establecer un equilibrio entre la consistencia teórica y la viabilidad. Además, se ha encontrado que el marco es aplicable a diversos dominios Eppler y Wittig, (2000). La estructura del marco es jerárquica, y organiza aspectos de calidad de datos en 15 dimensiones para comprender las cuatro categorías principales de calidad de datos Moges *et al.*, (2013). La Figura 4 proporciona una visión general de las dimensiones de calidad de datos consideradas en el estudio de Wang y Strong, (1996).

Figura 4. Esquema conceptual de calidad de datos
Fuente: (Wang y Strong, 1996)

Para identificar las dimensiones de calidad de datos más recurrentes, el presente trabajo sugiere en primer lugar adoptar el marco de calidad de datos propuesto por Wang y Strong (1996). Si este no se ajusta completamente a su contexto entonces valorar si se ajusta parcialmente y determinar otras dimensiones de calidad. Estas últimas pueden ser identificadas mediante la revisión bibliográfica de otras investigaciones con casos de estudio similares, o mediante el diseño y procesamiento estadístico de encuestas a los usuarios y expertos del negocio implicados en el contexto de aplicación.

Categorías intrínseca y contextual

La calidad de datos se puede medir a través de muchas dimensiones como la exactitud, la completitud, el tiempo de vida, la relevancia, la objetividad, la credibilidad, entre otras (Wang, 1998; Eppler y Wittig, 2000). Algunas de estas dimensiones se prestan para ser medidas haciendo uso de la medición objetiva, que es intrínseca a los datos en sí, independientemente del contexto en el que se utilicen, por ejemplo exactitud y objetividad. Sin embargo, existen dimensiones de calidad de datos que no se pueden medir objetivamente, por ejemplo las dimensiones relevancia y credibilidad (Fisher *et al., 2003;* Watts y Zhang, *2004*), las cuales tienden a variar con el contexto de uso. La relevancia de datos depende principalmente de la tarea que se desarrolle, ya que los datos que son altamente relevantes para una tarea pueden ser irrelevantes para otra; por ejemplo, diferentes datos son requeridos en tareas de ventas a la hora de realizar las hojas de balance, siendo estos datos irrelevantes para las tareas de *marketing*. Para entender los efectos contextuales de la calidad de datos, es importante tener en cuenta factores relacionados con el uso de los datos. En este sentido, varios factores como la relevancia de los datos en una tarea, la capacidad del usuario para entenderla, y la

claridad de la tarea, afectan la usabilidad de dichos datos (Watts *et al., 2009*). Desde esta perspectiva de uso, la evaluación de la calidad de datos tiende a ser contextual (Moges *et al.*, 2013).

Los investigadores Fisher *et al. (2003)* y Shankaranarayanan y Zhu (2012*)* identificaron el impacto de la experiencia de los usuarios (principiante o experto en el dominio), el tipo de tarea (simple o compleja) que estos llevan a cabo y las limitaciones de tiempo que presentan para desarrollarlas, como el posible uso de información relacionada con la calidad de datos. Sus resultados indican que cuando aumenta el nivel de experiencia de los usuarios y disminuye la complejidad de la tarea que realizan, se utiliza con más frecuencia la información sobre calidad de datos en tareas para la toma de decisiones. Asimismo, otros autores como Price y Shanks (2011) investigaron el impacto que poseen diferentes estrategias para la toma de decisiones sobre el uso de información relacionada con calidad de datos.

En general, todos estos estudios ilustran la existencia de factores que pueden afectar la evaluación de la calidad de datos y estimulan a investigar si otros factores, como la existencia de equipos de especialistas en calidad o el tamaño de las organizaciones, podrían impactar en dicha evaluación (Moges *et al.*, 2013).

Categorías representacional y accesibilidad

Las dimensiones de calidad de datos mencionadas con más frecuencia en las categorías de calidad de datos representacional y accesibilidad son: consistencia representacional, facilidad de comprensión, accesibilidad y seguridad. La consistencia representacional y la facilidad de comprensión evalúan la representación y la comprensión de datos respectivamente. Los problemas típicos como el uso de diferentes formas de difusión o propagación de datos, diferentes formatos y diferentes nombres para las columnas o filas similares, son tratados por la dimensión consistencia representacional. Por otro lado, las dos últimas dimensiones de calidad de datos evalúan la facilidad de acceso y la seguridad de los datos, respectivamente. La dimensión accesibilidad se refiere al grado en que los datos están disponibles o accesibles, por ejemplo, los datos pueden ser clasificados como inaccesibles si la brecha entre la entrada y el tiempo de entrega de salida es demasiado grande *(Strong et al., 1997*).

Dimensiones de calidad de datos más comunes

Existen discrepancias en la definición de la mayoría de las dimensiones de calidad de datos debido a la naturaleza contextual del término calidad. Además, no existe un acuerdo general que indique que un conjunto específico de dimensiones defina la

calidad de todos los datos, o en el que se defina el significado exacto de cada dimensión. No obstante, mediante el análisis de los estudios realizados por (Jarke *et al.*, 1995, Redman, 1996, Wand y Wang, 1996, Wang y Strong, 1996, Bovee *et al.*, 2001, Naumann, 2002) sobre este tema, es posible determinar un conjunto básico de las dimensiones de calidad de datos que se consideran como las más comunes. Estas dimensiones son: la exactitud, la completitud, la consistencia y el tiempo de vida del dato, las cuales constituyen el centro de atención de estos autores y generalmente son el factor común entre la mayoría de las investigaciones que se realizan para identificar las dimensiones de calidad de datos (Scannapieco y Catarci, 2002).

Exactitud

Varias definiciones son proporcionadas para el término exactitud. Los autores Wang y Strong (1996) definen exactitud como "el grado en que los datos son correctos, confiables y certificados". Ballou y Pazer (1985) especifican que los datos son exactos, cuando los valores de los datos almacenados en la base de datos se corresponden con los valores reales. Según Redman (1996), la exactitud se define como una medida de la proximidad de un valor de datos *v*, a algún otro valor *v´*, que se considera correcto. En esta definición, de forma general, dos tipos de exactitud se pueden distinguir: la sintáctica y la semántica. Las metodologías de calidad de datos sólo tienen en cuenta la exactitud sintáctica y la definen como la cercanía de un valor *v*, a los elementos del dominio de definición correspondiente *D*. En la exactitud sintáctica, no interesa la comparación de *v* con su valor en el mundo real *v´*; más bien, lo que se busca es comprobar si *v* es uno de los valores de *D*, o cuán cerca está de los valores en *D*. Por ejemplo, *v* = **"***Jean***"** se considera sintácticamente preciso incluso si *v'* = **"***John***"** (Batini *et al.*, 2009).

Completitud

La completitud es definida como el grado en el que un conjunto de datos dado, incluye datos que describen el conjunto correspondiente de objetos del mundo real (Batini *et al.*, 2009).

La Tabla 3 presenta contribuciones de varias investigaciones que proporcionan varias definiciones de la dimensión completitud. Al comparar estas definiciones, se puede observar que todas de una manera u otra coinciden en que la completitud tiene estrecha relación con la proporción que existe entre la cantidad de valores ausentes y los valores reales. A pesar de esto, las definiciones difieren en el contexto al que se refieren, por ejemplo sistemas de información en Wand y Wang (1996), almacenes de datos en (Jarke *et al.*, 1995) y entidades en Bovee *et al.*, (2001).

Tabla 3: Definiciones proporcionadas para la dimensión completitud

Referencia	**Definición**
(Wand y Wang, 1996)	Capacidad de un sistema de información para representar todos los estados significativos de un sistema del mundo real
(Wang y Strong, 1996)	Grado en el que los datos son de suficiente amplitud, profundidad y alcance para la tarea en cuestión
(Redman, 1996)	Grado en el cual son incluidos valores en una colección de datos
(Jarke *et al.*, 1995)	Porcentaje de información del mundo real entrado en fuentes de datos y/o almacenes de datos
(Bovee *et al.*, 2001)	Información en la que se tiene todas las partes requeridas de una descripción de entidades
(Naumann, 2002)	Relación entre el número de valores no nulos en una fuente y el tamaño de la relación universal
(Liu y Chi, 2002)	Todos los valores que se supone sean recopilados de acuerdo con una teoría de recopilación

Fuente: (Batini *et al.*, 2009)

En el área de investigación de bases de datos relacionales, la completitud a menudo se relaciona con el significado de los valores nulos. Un valor nulo tiene el significado general de un valor perdido, o sea un valor que existe en el mundo real, pero no está disponible en una colección de datos. Con el fin de caracterizar la completitud, es importante entender por qué el valor no se encuentra. Un valor puede faltar ya sea porque existe pero no se conoce, porque no existe, o porque no se sabe si existe (Atzeni y De Antonellis, 1993). A modo de ejemplo, en la Tabla 4 se muestra la tabla *Persona* con los atributos nombre, apellido, fecha de nacimiento y correo electrónico. Donde para las tuplas con id = 2, 3 y 4 el valor del correo electrónico es nulo, sin embargo no se consideran casos de incompletitud (Batini *et al.*, 2009).

Tabla 4: Valores nulos y datos completos

ID	Nombre	Apellido	Fecha de nacimiento	Correo electrónico
1	John	Smith	03/17/1974	smith@abc.it
2	Edward	Monroe	02/03/1967	Null
3	Anthony	White	01/01/1936	Null
4	Marianne	Collins	11/20/1995	Null

Fuente: (Batini *et al.*, 2009)

Consistencia

La dimensión consistencia se refiere a la violación de las reglas semánticas definidas sobre los elementos de un conjunto de datos. Con referencia a la teoría relacional, las restricciones de integridad son un tipo de dichas reglas semánticas. En el ámbito estadístico, las ediciones de datos son las reglas semánticas típicas que permiten comprobaciones de la consistencia de datos. En la teoría relacional se pueden distinguir dos categorías fundamentales para las restricciones de integridad nombradas: restricciones de intra-relación y de interrelación. Las restricciones de intra-relación definen el rango de valores admisibles para el dominio de un atributo. Ejemplos de esto son "la edad debe oscilar entre 0 y 120 años", o "si trabajo en años es inferior a 3, entonces salario no puede ser mayor de 25.000 pesos al año". Por otro lado, las restricciones de interrelación implican atributos de diferentes relaciones Batini *et al.*, (2009).

En la literatura existe una gran variedad de información sobre las bases de datos consistentes, vinculadas a la teoría relacional y a las restricciones de integridad. Por ejemplo, en Arenas *et al.*, (1999) toman en consideración el problema de la caracterización lógica de la noción de respuesta consistente en una base de datos relacional, el cual se puede violar dadas las restricciones de integridad. A razón de esto, los autores proponen un método para calcular respuestas consistentes demostrando su

solidez e integridad. Según Calı *et al.*, (2004), las restricciones de integridad también se han estudiado como facilitadoras de la integración de datos.

En el área estadística, los datos de los cuestionarios de un censo tienen una estructura que corresponde a la de un esquema de cuestionario. Las reglas semánticas, llamadas ediciones, pueden ser definidas en el esquema de cuestionario para especificar el conjunto correcto de respuestas. Tales reglas denotan típicamente condiciones de error. Por ejemplo una edición podría ser: si el estatus marital es casado, entonces la edad no debe ser inferior a 14 años. Después de la detección de registros erróneos el acto de restaurar los valores correctos se llama imputación (Fellegi y Holt, 1976).

Relacionadas con el tiempo

Un aspecto importante en los datos es su validez en el tiempo. Las principales dimensiones relacionadas con el tiempo que se proponen en la literatura son: actualidad, volatilidad y tiempo de vida. En la Tabla 5 se muestran las definiciones propuestas por varios autores para estas tres dimensiones de tiempo.

Tabla 5: Definiciones de dimensiones relacionadas con el tiempo

Referencia	Definición
(Wand y Wang, 1996)	El tiempo de vida se refiere solamente a la demora entre el cambio de un estado del mundo real y la resultante modificación del estado del sistema de información
(Wang y Strong, 1996)	El tiempo de vida es el grado en el que la edad de los datos es apropiada para una tarea en cuestión
(Redman, 1996)	La actualidad es el grado en el que un dato se encuentra actualizado. El valor de un dato se encuentra actualizado si este es correcto, a pesar de posibles discrepancias que puedan existir causadas por cambios de retrazos del tiempo al valor correcto
(Jarke *et al.*, 1995)	La actualidad describe el momento en el cual la información es introducida en las fuentes y/o almacenes de datos La volatilidad describe el período de tiempo durante el cual la información es válida en el mundo real
(Bovee *et al.*, 2001)	El tiempo de vida tiene dos componentes: la actualidad y la volatilidad La actualidad es una medida de cuan antigua es la información, basada en

	cuánto tiempo hace desde que esta fue grabada La volatilidad es una medida de la inestabilidad de la información, o sea, la frecuencia del cambio de un valor para un atributo de entidad
(Naumann, 2002)	El tiempo de vida es la edad promedio de los datos en una fuente
(Liu y Chi, 2002)	El tiempo de vida es el grado en el que los datos se encuentran suficientemente actualizados para una tarea

Fuente: Batini *et al.*, (2009)

El resumen de definiciones que se presenta en la Tabla 5 demuestra que no existe una definición única y exacta para las dimensiones relacionadas con el tiempo, típicamente la actualidad y el tiempo de vida se utilizan para referirse al mismo concepto. Se considera que la definición expresada por Bovee *et al.*, (2001) es la más completa porque expresa la actualidad y la volatilidad como los dos componentes principales del tiempo de vida de un dato, o sea destaca la estrecha relación que existe entre ellas. Aunque debe tenerse en cuenta que la mejor definición es aquella que se adecue más al contexto en el que se esté trabajando.

Métricas de calidad de datos

Cada dimensión de calidad de datos identificada posee una o varias métricas (formas de evaluación) asociadas, las cuales se definen en la fase de medición de la metodología TDQM.

El nivel de calidad de datos en las organizaciones puede ser evaluado mediante el uso de cuestionarios o métricas Moges *et al.*, (2013). Las métricas de calidad de datos son una forma de calcular "las puntuaciones de calidad", además brindan una medida sobre el nivel de calidad que poseen los datos. Las métricas de calidad de datos se suelen utilizar para definir los criterios de aceptabilidad de los datos y para realizar la corrección y mejora de los mismos Emran *et al.*, (2012).

En Pipino *et al.* (2002) se desarrollaron métricas de calidad de datos basados en tres formas funcionales: la relación simple, la operación de min/max y la media ponderada.

La primera de estas mide la relación existente entre los resultados deseados con los resultados totales. A pesar de que la mayoría de las personas tiende a medir las excepciones o errores, la forma preferida para realizar la medición es el número de resultados no deseados dividido por el total de resultados y restado con 1. Esta relación simple se adhiere a la convención de que 1 representa el resultado más deseable y 0 el resultado menos deseable (Ballou *et al.*, 1998; Ballou y Pazer, 1985; Huang *et al.*, 1998;

Redman, 1996). Aunque una relación que ilustre los resultados indeseables brinda la misma información que una donde se ilustren los resultados más deseables, Pipino *et al.* (2002) sugieren manejar preferiblemente la relación que muestra resultados positivos, ya que esta forma es útil para comparaciones longitudinales que ilustran las tendencias de mejora continua. También plantean que muchas dimensiones tradicionales de calidad de datos tales como la exactitud, la completitud y la consistencia toman esta forma de medición.

En caso de medir variables individuales es usual utilizar una relación simple. Sin embargo, para manejar dimensiones que requieran la agregación de múltiples indicadores de calidad de datos (variables) se sugiere aplicar la operación de mínimo o de máximo, donde se calcula el valor mínimo (o máximo) entre los valores normalizados de los indicadores individuales de calidad de datos. El operador mínimo (min) es conservador en el sentido de que asigna a la dimensión un valor total no mayor que el valor de su indicador de calidad de datos más débil, evaluado y normalizado entre los valores 0 y 1 (Pipino *et al.*, 2002). Dos ejemplos interesantes de dimensiones que pueden hacer uso del operador mínimo son la credibilidad y la cantidad adecuada de los datos. El operador máximo resulta útil en métricas más complejas aplicables a las dimensiones del tiempo de vida y la accesibilidad (Pipino *et al.*, 2002).

Para el caso multivariable, una alternativa al operador mínimo es el uso de la media ponderada de variables. Por ejemplo, si una empresa tiene una buena comprensión de la importancia de cada variable para la evaluación global de una dimensión, entonces el uso de una media ponderada de las variables es lo adecuado. Para asegurar la calificación esta es normalizada, cada factor de ponderación debe estar entre los valores 0 y 1 y la suma de todos los factores de ponderación debe ser igual a 1 (Pipino *et al.*, 2002).

Las métricas pueden ser formuladas para tareas independientes o tareas dependientes (Moges *et al.*, 2013). Las métricas para tareas independientes reflejan estados de los datos sin el conocimiento del contexto de su aplicación y pueden aplicarse a cualquier conjunto de datos, independientemente de la tarea en cuestión. Las métricas para las tareas dependientes se desarrollan en contextos de aplicación específicos e incluyen reglas de organización de negocios, regulaciones de compañías y del gobierno y restricciones establecidas por administradores de base de datos (Pipino *et al.*, 2002).

La clave para la medición es el desarrollo de métricas de calidad de datos, las cuales proporcionan las medidas básicas de calidad tales como la exactitud, el tiempo de vida, la completitud y la consistencia de los datos (Ballou y Pazer, 1985; Wang y Lee, 1998). Por ejemplo, en la base de datos “cuenta del cliente”, las métricas de calidad de datos pueden ser diseñadas para realizar un seguimiento a los datos (Wang, 1998):

- El porcentaje de los códigos postales incorrectos de direcciones de clientes, que se encuentran en una cuenta de cliente seleccionada al azar (libre de error).
- Un indicador de cuando los datos de la cuenta del cliente se actualizaron por última vez (el tiempo de vida o la actualidad para la comercialización de bases de datos y fines reglamentarios).
- El porcentaje de cuentas inexistentes o el número de cuentas con valores ausentes en el campo de código de la industria (completitud).
- El número de registros que violan la integridad referencial (consistencia).

Métricas para dimensiones de calidad de datos más comunes

Una cantidad considerable de métricas han sido y están siendo desarrolladas por diferentes investigadores y organizaciones de calidad de datos. El proceso de elegir las variables o componentes específicas para la medición puede ser más difícil que definir la métrica en general. A pesar de que cada métrica cae dentro de una categoría dimensional específica, la medida para evaluar una dimensión determinada puede variar de una organización a otra (Lee *et al.*, 2006).

En la Tabla 6 se muestra un resumen de las métricas propuestas por diferentes autores, para las dimensiones de calidad de datos más comunes.

Tabla 6: Métricas de calidad de datos de la literatura Basado en:

Dimensión	Métrica de calidad de datos (Tarea independiente)	Referencia
		(Pipino *et al.*, 2002; Dani *et al.*, 2010)
		(Fisher *et al.*, 2009)
Exactitud		(Sessions y Valtorta, 2009)
		(Pipino *et al.*, 2002; Batini *et al.*, 2009)
		(Codd, 1982; Pipino *et al.*, 2002)

Completitud		(Batini *et al.*, 2009)
		(Pipino *et al.*, 2002)
Consistencia		(Batini *et al.*, 2009)
Actualidad		(Pipino *et al.*, 2002, Batini *et al.*, 2009; Moges *et al.*, 2013)
Volatilidad		(Pipino *et al.*, 2002; Batini *et al.*, 2009; Moges *et al.*, 2013)
Tiempo de vida		(Pipino *et al.*, 2002; Dani *et al.*, 2010)
Dimensión	**Métrica de calidad de datos (Tarea dependiente)**	**Referencia**
Tiempo de vida		(Pipino *et al.*, 2002; Batini *et al.*, 2009; Moges *et al.*, 2013)

(Moges *et al.*, 2013)

Algunas connotaciones importantes para las tareas independientes no mencionadas con anterioridad a tener en cuenta para un mejor entendimiento de la Tabla 6 son:

- En el caso de la exactitud,
    - *f* indica "función de".
- En el caso del tiempo de vida,
    - *Qact* es el nivel de actualidad de los datos,
    - *A* es un atributo,
    - *W* es un valor de atributo,
    - *edad* se refiere a la diferencia entre el momento en que la calidad de datos es evaluada y el momento de adquisición de los datos y

- *disminución* se refiere a la velocidad promedio de descenso del tiempo de vida de los valores del atributo, para el atributo bajo consideración (Moges *et al.*, 2013).

Cualquiera sea la naturaleza de las métricas de calidad de datos, estas son implementadas como parte de un nuevo sistema de fabricación de información o como un complemento de rutinas de utilidad en un sistema existente. Con las métricas de calidad de datos, las medidas pueden ser obtenidas a lo largo de varias dimensiones para el análisis (Pipino *et al.*, 2002). A continuación se profundiza en la forma de medición de las dimensiones de calidad de datos más comunes.

La dimensión *completitud* se puede analizar desde diversos puntos de vista, lo que conduce a la aparición de diferentes métricas. En el nivel más abstracto, se puede definir el concepto de completitud de esquema como el grado en el cual las entidades y atributos se encuentran presentes en el esquema. En el nivel de datos, se puede definir la completitud de columnas como una función que evalúa los valores perdidos en una columna de una tabla. Esta medición corresponde a la integridad de columnas de Codd's (1982), que evalúa los valores perdidos. Un tercer tipo se denomina completitud poblacional, donde si una columna debe contener al menos una ocurrencia de 50 estados, pero sólo contiene 43 de estos, entonces tenemos incompletitud poblacional. Cada uno de los tres tipos (completitud de esquema, completitud de columnas, y completitud poblacional) pueden ser medidos tomando la relación de la cantidad de ítems incompletos entre el número total de elementos y substraerle 1 a esta división (Pipino *et al.*, 2002).

La dimensión *consistencia* se puede ver como la consistencia de valores redundantes a través de las tablas. Las restricciones de integridad referencial de Codd's son una instancia de este tipo de consistencia. Al igual que las dimensiones discutidas previamente, una métrica para medir la consistencia puede ser la relación de violaciones de un tipo de consistencia específica entre el número total de comprobaciones de consistencias menos uno (Pipino *et al.*, 2002).

Ballou *et al.* (1998) sugieren medir el *tiempo de vida* como el máximo entre dos términos: cero y 1 menos la relación de la actualidad entre la volatilidad, como se muestra en la Tabla 6. Esta métrica depende del contexto en el que se vaya a aplicar. Si la dimensión tiempo de vida no es crítica para la tarea en cuestión, entonces una medida de sensibilidad más liberal se puede aplicar. Inversamente si la dimensión es muy crítica, una medida de sensibilidad conservadora se sugiere (Moges *et al.*, 2013). Además, se define la *actualidad* como la edad más el tiempo de entrega menos el tiempo de entrada. También se refiere a la *volatilidad* como la longitud del tiempo de los datos que sigue siendo válida, el tiempo de entrega se refiere a cuando los datos son

entregados al usuario, el tiempo de entrada se refiere a cuando se reciben los datos por el sistema y la edad se refiere a la edad de los datos cuando se reciben primeramente por el sistema (Pipino *et al.*, 2002). En dicha métrica se utiliza un exponente como un factor de sensibilidad, elevando el valor máximo a este exponente, ver Tabla 6. El valor del exponente es una tarea dependiente y refleja el juicio del analista. Por ejemplo, supongamos que la calificación del tiempo de vida sin utilizar el factor de sensibilidad (equivalente a un factor de sensibilidad igual a 1) es 0,81. El uso de un factor de sensibilidad igual a 2 debería entonces producir una calificación para el tiempo de vida de 0,64 (un factor de mayor sensibilidad refleja el hecho de que los datos se conviertan más rápidos en menos tiempo) y 0,9 cuando el factor de sensibilidad es 0,5 (un factor de menor sensibilidad refleja el hecho de que los datos pierden el tiempo de vida a un menor ritmo) (Pipino *et al.*, 2002).

Más reciente, en Fisher *et al.*, (2009) proponen una métrica para la *exactitud* cambiando la escala de razón simple a un vector de aproximación, que incluye porcentajes, una medida de aleatoriedad, y una distribución de probabilidad. La métrica combina una relación simple, mediante el cálculo del número de celdas con errores entre el número total de celdas, con una medida de aleatoriedad, calculada mediante el algoritmo Lempel-Ziv[1] (L-Z, por sus siglas en inglés) para medidas de complejidad. Este algoritmo se utiliza para diferenciar si los errores en una base de datos son aleatorios o sistemáticos en la naturaleza. Una vez determinada la aleatoriedad de los errores, una distribución de probabilidad se utiliza para ayudar a abordar diversas cuestiones de gestión. La métrica se basa en la suposición de que el valor corresponde a la gama de validez posible (Moges *et al.*, 2013). Del mismo modo, (Sessions y Valtorta, 2009) sugieren un enfoque de medición para la exactitud usando redes bayesianas[2], véase la Tabla 6.

Perfilado de datos

Una vez terminada la fase de medición de la calidad de datos, con los resultados obtenidos de dicha fase se investigan las causas que conllevan a los problemas de calidad de datos detectados, este proceso se realiza durante la fase de análisis de la metodología TDQM.

Investigar cuáles son las características de los datos y qué provecho se puede extraer de ellos se ha convertido en un objetivo prioritario en empresas e instituciones que manejan grandes volúmenes de información y se enfrentan a fuertes competencias de mercado. En este marco adquiere especial atención los procesos de perfilado de datos, los cuales mediante la utilización de métodos estadísticos y de minería de datos brindan información detallada y descriptiva sobre las fuentes de datos (González, 2010).

El perfilado de datos se puede definir como el proceso de análisis de las fuentes de datos con respecto al dominio[3] de calidad de datos (Marshall, 2007) para identificar los problemas de calidad que puedan presentar los datos.

Desde el punto de vista del analista, se consideran como aspectos de suma importancia para entender los problemas de calidad de datos más comunes: determinar el alcance de las causas subyacentes principales de los problemas de calidad de datos y planificar el diseño de las herramientas que se pueden utilizar para hacer frente a dichos problemas, de tal manera que la clasificación formada sea útil para las comunidades de almacenes de datos y calidad de datos (Pandey, 2014). Es crucial la identificación de los problemas y las causas que conllevan a los mismos para asegurar la entrega de información precisa antes de iniciar un plan de limpieza de datos.

Causas y problemas que conllevan a una mala calidad de datos

La taxonomía que se presenta a continuación para los problemas de calidad que pueden presentar los datos se basa en lo expresado por Kim *et al.*, (2003), Müller y Freytag (2005), Oliveira *et al.*, (2005), Rahm y Do, (2000). En esta taxonomía se dividen los problemas de calidad de datos a nivel de esquema y de instancia. Los problemas en el nivel de esquema pueden evitarse con un diseño de esquema mejorado y no dependen de los contenidos reales de los datos. Por otro lado, los problemas en el nivel de instancia están relacionados con el contenido de los datos, solo que estos no pueden ser evitados con una mejor definición de esquema, como en los lenguajes de definición de esquema, ya que estos no son lo suficientemente potentes para especificar todas las limitaciones de los datos requeridos. En esta taxonomía, los problemas con los datos en el nivel de instancia incluyen todos los problemas de calidad de datos que no se encuentran en el nivel de esquema, por lo que la taxonomía es completa en este punto (Barateiro y Galhardas, 2005).

Problemas de calidad de datos a nivel de esquema

Los problemas de calidad de datos en el nivel de esquema se pueden prevenir con un diseño de esquema mejorado y una correcta traducción del esquema e integración de todos los datos manejados. Los sistemas gestores de bases de datos relacionales (RDBMS, por sus siglas en inglés) proporcionan mecanismos importantes para asegurar definiciones de esquema. Por tanto, se pueden distinguir entre problemas de calidad de datos a nivel de esquema que pueden ser evitados mediante un RDBMS y problemas de calidad de datos a nivel de esquema que no pueden ser evitados mediante un RDBMS (Barateiro y Galhardas, 2005).

1. Evitados por los RDBMS

Durante la fase de diseño de bases de datos las restricciones de integridad pueden ser definidas para especificar diferentes condiciones que deben satisfacer los objetos en la base de datos (Türker y Gertz, 2001). En este sentido, SQL (ISO y ANSI, 1999) proporciona un lenguaje para apoyar la especificación de las siguientes restricciones de integridad declarativas:

- restricción no nulo, para evitar que una columna tome un valor nulo;
- restricción predeterminada, para especificar el valor predeterminado de una columna dada;
- restricción único, para definir que una columna o un conjunto de columnas deben tener valores únicos dentro de una tabla;
- restricción de clave principal, que corresponde a la restricción no nulo y a la restricción único;
- restricción de integridad referencial, para especificar atributos cuyos valores deben coincidir con los valores de una clave principal (o única) de una tabla externa;
- restricción de verificación, para especificar una condición definida por el usuario;
- restricción de dominio, para definir valores restringidos del dominio de columna y
- restricción de afirmación, define una restricción de integridad tabla-independiente. SQL (1999) también ofrece una definición procedural para las restricciones de integridad usando disparadores (*triggers*), que son procedimientos invocados por el RDBMS en respuesta a eventos específicos de bases de datos (Barateiro y Galhardas, 2005).

La siguiente lista resume los problemas de calidad de datos que pueden ser evitados con una definición adecuada de las restricciones de integridad (Barateiro y Galhardas, 2005).

- Datos perdidos: datos que no han sido llenados. Una restricción no nula puede evitar este problema.
- Tipo de datos incorrecto: violación del tipo de datos de una restricción; por ejemplo, la edad del empleado es “xy”. Las restricciones de dominio pueden evitar este problema.

- Valor de datos incorrecto: violación del rango de datos de una restricción; por ejemplo, si la edad de un empleado debe pertenecer al rango [18, 65], entonces una edad de 15 años es un valor de datos incorrecto. Las restricciones de verificación y dominio se utilizan para evitar este problema.
- Datos colgantes: los datos que en una tabla no tienen equivalente en otra tabla; por ejemplo, un identificador de departamento que no exista en una tabla Departamento, y sin embargo hay una referencia a este valor en una tabla Empleado. Este problema se aborda mediante restricciones de integridad referencial (es decir, claves o llaves foráneas).
- Duplicado exacto de datos: los diferentes registros tienen el mismo valor en un campo (o una combinación de campos) para los cuales la duplicación de valores no está permitida; por ejemplo, un número de seguro social de un empleado. Las restricciones de claves únicas y primarias pueden evitar duplicados exactos.
- Restricciones genéricas de dominio: registros o valores que violan una restricción de dominio de un atributo; por ejemplo, dependencias de atributos o un número máximo predefinido de filas. Las restricciones de afirmación y de dominio pueden evitar este problema. Sin embargo, estas características no son proporcionadas por la mayoría de los RDBMS (por ejemplo, Oracle 9i) que manejan este problema de datos utilizando los *triggers*.

2. No evitados por los RDBMS

Los siguientes problemas de calidad de datos no pueden ser manejados por las restricciones de integridad de los RDBMS (Barateiro y Galhardas, 2005):

- Datos categóricamente incorrectos: un valor de categoría que está fuera del rango de dicha categoría, por ejemplo, países y estados respectivos. El uso de un nivel de abstracción mal, por ejemplo, *"alimento congelado"* o "*pizza congelada"* en lugar de *"alimentos"* también se considera un tipo de datos incorrecto categóricamente.
- Datos desfasados temporalmente: los datos que violan una restricción temporal, que especifica el instante de tiempo o intervalo en el que los datos son válidos, por ejemplo, el salario de un empleado ya no es válido cuando este empleado es ascendido.
- Datos espaciales inconsistentes: inconsistencias entre los datos espaciales (por ejemplo, las coordenadas, las formas) cuando están almacenados en múltiples campos, por ejemplo, las coordenadas rectangulares de un punto en una tabla deben combinarse para producir un rectángulo cerrado.

- Conflictos de nombre: el mismo nombre de campo se utiliza para diferentes objetos (homónimos), o se utilizan nombres diferentes para el mismo objeto (sinónimos).
- Conflictos estructurales: diferentes representaciones de esquema del mismo objeto en diferentes tablas o bases de datos, por ejemplo, una dirección se puede representar en un campo de forma libre o descomponerse en los campos calle, ciudad, estado, etc.

Aunque, en el nombre y en las estructuras pueden producirse conflictos dentro de un mismo esquema de datos, con frecuencia surgen en un escenario multi-esquema (Barateiro y Galhardas, 2005).

Problemas de calidad de datos a nivel de instancia

Los problemas con los datos en el nivel de instancia se refieren a errores e inconsistencias en los datos que no son visibles o evitados a nivel de esquema. Se debe tener en cuenta que las instancias de datos también reflejan los problemas a nivel de esquema (por ejemplo, un registro con un valor nulo en un campo requerido). Los problemas de datos a nivel de instancia se pueden dividir en problemas de un único registro y problemas de múltiples registros (Barateiro y Galhardas, 2005).

1. Registro simple

Los problemas con los datos en un registro individual o simple se refieren a uno o varios atributos de un único registro. En otras palabras, este tipo de problemas se relaciona con una sola entidad y no depende de otra información almacenada en la base de datos (Barateiro y Galhardas, 2005). Los problemas son:

- Falta de datos en un campo no nulo: atributos que son rellenados con algún valor ficticio; por ejemplo, un número de seguro social 99999 es un valor indefinido utilizado para superar la restricción no nula.
- Datos erróneos: datos que son válidos, pero no se ajustan a la entidad real; un ejemplo de datos erróneos es 36 que indica la edad de un empleado cuando este en realidad tiene 35 años.
- Errores ortográficos: palabras mal escritas en campos de bases de datos; por ejemplo, *"Jhon Stevens"* en lugar de *"John Stevens"*.
- Valores implícitos: existencia de datos extraños en algún campo de datos; un ejemplo común de valores implícitos es la inserción de un título en un campo de nombre, por ejemplo, *"Presidente John Stevens"*.
- Valores de campos perdidos: datos que son almacenados en el campo equivocado. Por ejemplo, el valor *"Cuba"* ubicado en el atributo ciudad.

- Datos ambiguos: datos que pueden ser interpretados en más de una manera, con diferentes significados. La existencia de datos ambiguos puede ocurrir debido a la presencia de abreviaturas o contextos incompletos, como los siguientes:
  - Abreviaturas: el nombre abreviado *"J. Stevens"* puede ser ampliado de diferentes formas, como: *"John Stevens"*, *"Jack Stevens"*, *"Jeff Stevens"*, etc.
  - Contextos incompletos: el nombre de la ciudad *"Miami"* puede ser soportado para el estado de la Florida, o el estado de Ohio.

2. Registros múltiples

Los problemas con los datos en múltiples registros no pueden ser detectados considerando cada registro por separado, ya que como el nombre lo indica, se está haciendo referencia a más de un registro. Se debe tener en cuenta que pueden ocurrir múltiples problemas de registro entre los registros que pertenecen al mismo conjunto de entidades (tablas) o a diferentes conjuntos de entidades (correspondientes a diferentes tablas o incluso a diferentes bases de datos) (Barateiro y Galhardas, 2005):

- Registros duplicados: son aquellos que representan la misma entidad real y no contienen información contradictoria; por ejemplo, los siguientes registros de empleados se consideran duplicados: Empleado1 (Nombre = *"John Stevens"*, Dirección = *"223, Primera Avenida, Ciudad de Nueva York"*, Nacimiento = 01/01/1975); Empleado 2 (Nombre = *"J. Stevens"*, Dirección = *"223, 1st Avenue, New York"*, Nacimiento = 01/01/1975).
- Contradicción de registros: existe entre registros que representan la misma entidad real y contengan algún tipo de información contradictoria; por ejemplo, si se consideran los registros Empleado 1 y Empleado 2, pero con la siguiente información: Empleado1 (Nombre = *"John Stevens"*, Dirección = *"223, Primera Avenida, Ciudad de Nueva York"*, nacimiento = 01/01/1975); Empleado2 (Nombre = *"John Stevens"*, Dirección = *"223, Primera Avenida, Ciudad de Nueva York"*, nacimiento = 01/01/1965).
- Datos no estandarizados: diferentes registros que no utilizan las mismas representaciones de datos, invalidando así su comparación:
  - Transposición de palabras: en un solo campo, las palabras pueden aparecer con diferentes ordenamientos; por ejemplo, los nombres *"John Stevens"* y *"Smith, Jack"* no utilizan la misma regla de ordenación.
  - Formato de codificación: uso de diferentes formatos de codificación, por ejemplo, ASCII, UTF-8.

- Formato de representación: diferentes representaciones de un formato para la misma información; un ejemplo es el formato de moneda que puede ser € 10.5, 10.5 €, etc.
- Unidad de medida: diferentes unidades utilizadas en registros distintos; por ejemplo, las distancias dadas en cm y pulgadas.

Algoritmos o métodos empleados en el perfilado de datos

Como se mencionó anteriormente, uno de los inconvenientes que pueden presentar los datos almacenados es cuando una misma entidad del mundo real se almacena más de una vez de diferentes formas. El proceso para detectar este tipo de problemas se conoce como registro de enlace (*record linkage)* en el área de la estadística, radicalización de base de datos (*database hardening)* en el área de la inteligencia artificial y mezclar-depurar (*merge-purge)*, des-duplicar datos (*data deduplication)* o identificación de instancias (*instance identification)* en el área de las bases de datos. Otros nombres como resolución correferencial (*coreference resolution)* y detección de registros duplicados (*duplicate record detection)* también se usan con frecuencia (Amón *et al.*, 2012).

Este proceso fue identificado inicialmente por Dunn (1946). Algunos fundamentos probabilísticos fueron desarrollados posteriormente por Newcombe y Kennedy (1962), y formalizados por Fellegi y Holt (1976) como una regla de decisión probabilística. Algunas mejoras de este modelo han sido propuestas por Winkler y Census (1993). Esencialmente, el proceso es como sigue: dado un conjunto *R* de registros: (i) se define un umbral en el intervalo cerrado [0,1], (ii) se compara cada registro con los demás registros y (iii) si la similitud entre una pareja de registros es mayor o igual que el umbral, se supone que son duplicados y se considera que son representaciones de una misma entidad del mundo real (Amón *et al.*, 2012).

A consecuencia de esto se han desarrollado funciones de similitud para la detección de duplicados, algunas de ellas son de tipo fonético como *Soundex* (Raghavan y Allan, 2004), *Metaphone* (Philips, 1990), *Double Metaphone* (Philips, 2000), *Onca* (Gill, 1997) y *Nysiis* (Taft, 1970). Estas técnicas se basan en la forma en que se pronuncian las palabras en un idioma en particular y no están orientadas al idioma español. Otro tipo de funciones de similitud se basan en emparejamiento de patrones. En esta última categoría, se encuentran técnicas como *Levenshtein*, *Brecha Afín*, *Smith-Waterman*, *Jaro*, *Jaro-Winkler*, *Bi-grams*, *Tri -grams*, *Monge-Elkan* y *SoftTF-IDF* (Amón *et al.*, 2012).

En los gestores de datos el intento más generalizado de buscar cadenas similares (no exactas) está relacionado con la utilización de la función *Soundex* (cadena) (Gálvez, 2006). *Soundex* es un algoritmo de codificación fonética, que convierte una palabra en

un código (Else, 2002). El código *Soundex* consiste en sustituir las consonantes de la palabra afectada por un número, si es necesario se agregan ceros al final del código para conformar un código de cuatro dígitos. *Soundex* elige la clasificación de los caracteres con base en el lugar de articulación de la lengua inglesa. La Tabla 7 presenta las equivalencias usadas por *Soundex* (Amón *et al.*, 2012).

Tabla 7: Equivalencias *Soundex*

Dígito	**Caracteres**
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

*Fuente: (*Amón *et al., 2012)*

Debido a que en el idioma inglés las letras *A*, *E*, *I*, *O*, *U*, *H*, *W* e *Y* no hacen diferenciación fonética, estas son descartadas. Adicionalmente existen otras reglas complementarias para la codificación de letras dobles (si el texto tiene letras dobles, estas se deben tratar como una sola) y para letras con el mismo código, las cuales al realizar la operación de descarte quedan una al lado de la otra (si el texto tiene diferentes letras una al lado de la otra, pero tienen el mismo número en la guía de codificación *Soundex*, estas se deben tratar como una sola letra), entre otras. Por ejemplo: Giraldo se codifica *G643* (G, 6 por la R, 4 por la L, 3 por la D, las otras letras se descartan). Juan se codifica *J500* (J, 5 por la N, las otras letras se descartan, y se agregan dos ceros) (Amón *et al.*, 2012).

Las limitaciones del algoritmo *Soundex* han sido documentadas en (Patman y Shaefer (2003), y Stanier (1990) y han dado lugar a varias mejoras, pero ninguna orientada hacia el idioma español. Es claro, que *Soundex* no está orientado al español ya que ni siquiera contempla el juego de caracteres (*"ñ", "ll"*). Asimismo, la dependencia de la letra inicial, la agrupación por punto de articulación del idioma inglés y el estar limitado a cuatro caracteres implica que no es eficiente para detectar errores ortográficos comunes en el idioma español (Amón *et al.*, 2012).

En el caso del emparejamiento de patrones, la distancia de edición entre dos cadenas fue introducida por Levenshtein (1965) como la cantidad mínima de operaciones de

inserción, eliminación y sustitución que hay que hacer para transformar una cadena en la otra. La función así definida se demuestra que constituye una métrica y ha tenido algunas variaciones en el tiempo. Levenshtein consideró todas las operaciones con costo unitario, pero en trabajos posteriores se planteó que era posible que cada una de las operaciones de edición tuvieran costos diferentes (Hall y Dowling, 1980), por lo que en Porrero, (2011) se adicionó una nueva operación: la transposición, intercambio de caracteres adyacentes que es un error frecuente en las personas que teclean rápido Zobel y Dart, (1996). No se han encontrado en la literatura heurísticas para determinar los costos de las operaciones, aunque en Scherbina, (2005); Smith y Waterman, 1981) se señala que a partir de experimentos se obtienen buenos resultados cuando los costos no son unitarios sino cuando tienen valor 2.

La métrica de *Smith Waterman* (1981), utilizada inicialmente para buscar alineación entre moléculas, utiliza la idea de la distancia de edición. Además, propone utilizar costos negativos en las operaciones de inserción y eliminación y costos positivos para las operaciones de sustitución y cuando haya coincidencia. O sea, se toma el costo de las operaciones de inserción y de eliminación en función del tamaño del *"hueco"* que se produce al hacer una de estas operaciones.

La métrica de *Jaro* (Jebamalar-Tamilselvi y Saravanan, 2008) es usada comúnmente para buscar similitudes entre nombres y en sistemas de enlaces de registros (Winkler, 2006). Se basa en el conteo de los caracteres comunes de ambas cadenas, el número de transposiciones y el empleo de una expresión donde intervienen estos valores y la longitud de las cadenas.

La métrica *Q-Gram* (Ukkonen, 1992) y algunas de sus variantes (Li *et al.*, 2007) se basan en la determinación de subcadenas de longitud q (habitualmente se usa *q* igual a 1, 2 o 3). Se buscan las subcadenas comunes en las dos cadenas pues se plantea que, si dos cadenas son similares, comparten un alto número de *Q-Gram*.

Las métricas basadas en *tokens* dividen las cadenas en partes (*tokens*), a partir de espacios en blanco, signos de puntuación, etc. Entre las métricas basadas en *tokens* se encuentra la similitud de *Monge-Elkan* (Monge y Elkan, 1996), que se calcula como la similitud máxima promedio entre una pareja de *tokens*. Una variante se presenta en Gelbukh *et al.*, (2009) donde se utiliza la media aritmética generalizada en lugar del promedio. Otra métrica basada en *tokens* es la *similaridad del coseno* (Cohen *et al.*, 2003), en la que a cada palabra se le asigna un peso en dependencia de la frecuencia relativa de aparición en la cadena, formándose vectores con estos valores y buscándose el coseno del ángulo que existe entre estos vectores. Estas técnicas son muy utilizadas en la recuperación de la información (Rema *et al.*, 2008; Chaudhuri *et al.*, 2009).

Según Porrero (2011), estas distancias, incluyendo las de edición, no tienen en cuenta de manera explícita los errores tipográficos. Considerando que la operación más frecuente dentro de los errores tipográficos es la sustitución, pudiera mejorarse el proceso de estandarización de cadenas de caracteres teniendo en cuenta la posición de los caracteres en el teclado al calcular la distancia entre dos cadenas.

Herramientas comúnmente usadas para el perfilado de datos

Las herramientas de calidad de datos tienen como objetivo detectar y corregir problemas de datos que afectan la precisión y la eficiencia de las aplicaciones de análisis de datos (Barateiro and Galhardas, 2005). Es comúnmente aceptado que las herramientas de calidad de datos se pueden agrupar de acuerdo con la parte del proceso de calidad de datos que estas cubren (Olson, 2003). En la actualidad se pueden utilizar herramientas que traen implementadas técnicas para el perfilado de datos de manera automatizada, las cuales se clasifican como herramientas comerciales y de investigación (Barateiro and Galhardas, 2005). Dentro de las herramientas comerciales se encuentran: *dfPower, ETLQ, Trillium, Migration Architect, WizWhy, WizRule, DataStage, Firstlogic, Hummingbird ETL, Informatica ETL, Sagent, SQL Server 2000 DTS, SQL Server 2005, SQL Server 2008, Data Cleaner, Oracle, Oracle 10g, Profiler, Sunopsis, Talend* e *IBM Ascential*. Dentro de las herramientas de investigación se encuentran *Potter´s Wheel, Ken State University Tool*, entre otras (Barateiro and Galhardas, 2005).

Tomando en cuenta las herramientas anteriormente mencionadas, se muestra en la Tabla 8 un resumen de los principales rasgos con los que cuentan las herramientas de calidad de datos, analizando diferentes funcionalidades que poseen las mismas. Donde el valor Y significa soportado, X no soportado, - información desconocida, N nativo, DB bases de datos relacionales, FF archivos planos, G gráfica; NG no gráfica y M significa manual respectivamente, ver Tabla 8.

Tabla 8: Funcionalidades de las herramientas usadas en el perfilado de datos

Herramienta	Fuentes de datos	Capacidad de extracción	Capacidad de carga	Actualizaciones incrementales	Interfaz	Repositorio de metadatos	Técnicas de rendimiento	Versiones	Biblioteca de funciones	Lenguaje vinculante	Depuración	Manejo de errores	Linaje de datos
dfPower	Varias	Y	Y	-	G	Y	Y	-	-	-	-	-	-
ETLQ	Varias	Y	Y	-	G	Y	Y	Y	Y	N	-	-	-
Migration Architect	Varias	-	-	-	G	Y	-	-	-	-	-	-	-
Trillium	Varias	Y	Y	-	G	Y	Y	Y	Y	N	Y	-	Y

WizWhy	DB,FF	-	-	-	G	-	Y	-	-	-	-	-	-
WizRule	DB,FF	-	-	-	G	-	Y	-	-	-	-	-	-
SQL Server 2008	Varias	Y	Y	-	G	-	-	-	Y	N	-	-	-
DataStage	Varias	Y	Y	-	G	Y	Y	Y	Y	Y	Y	Y	Y
Firstlogic	DB,FF	Y	Y	-	G	Y	Y	-	Y	Y	-	-	-
Hummingbird ETL	Varias	Y	Y	Y	G	Y	Y	Y	Y	N	Y	M	Y
Informatica ETL	Varias	Y	Y	Y	G	Y	Y	Y	Y	Y	Y	Y	Y
Sagent	Varias	Y	Y	X	G	Y	Y	X	Y	N, SQL	-	-	-
SQL Server 2000 DTS	Varias	Y	Y	X	G	X	-	-	Y	N	X	X	X
SQL Server 2005	Varias	Y	Y	-	G	-	-	-	Y	N	-	-	-
Oracle	-	-	-	-	G	-	-	-	-	-	-	-	-
Oracle 10g	-	-	-	-	G	-	-	-	-	-	-	-	-
DataCleaner	-	-	-	-	G	-	-	-	-	-	-	-	-
Profiler	-	-	-	-	G	-	-	-	-	-	-	-	-
Sunopsis	DB,FF	Y	Y	Y	G	Y	Y	Y	X	SQL	Y	Y	X
Talend	-	-	-	-	G	-	-	-	-	-	-	-	-
IBM Ascential	-	-	-	-	G	-	-	-	-	-	-	-	-
Potter´s Wheels	Y	-	ODBC	-	G	Y	-	-	-	-	Y	-	-
Ken State University Tool	-	-	X	-	NG	-	-	-	-	-	-	Y	-

Basado en: (Barateiro y Galhardas, 2005)

A continuación se explican los principales rasgos de las herramientas de calidad de datos que se muestran en la Tabla 8 (Barateiro y Galhardas, 2005):

- **Fuentes de datos:** las fuentes de datos pueden ser heterogéneas, por ejemplo bases de datos relacionales, archivos planos, archivos XML, hojas de cálculo, sistemas legados, paquetes de aplicaciones (como SAP), fuentes basadas en internet, software EAI (*Enterprise Application Integration*), entre otras.
- **Capacidad de extracción:** el proceso de extracción de datos desde las fuentes debe proporcionar las siguientes capacidades: (i) la posibilidad de programar las extracciones por hora, intervalo o evento, (ii) un conjunto de reglas para

seleccionar datos de la fuente y (iii) la capacidad para seleccionar y combinar registros de múltiples fuentes.

- **Capacidad de carga:** el proceso de carga de datos dentro del sistema destino debe ser capaz de: (i) cargar datos dentro de múltiples tipos de sistemas destinos, (ii) cargar datos dentro de los sistemas destinos heterogéneos en paralelo, (iii) actualizar y añadir datos dentro de las fuentes de datos del destino y (iv) crear automáticamente las tablas del destino.
- **Actualizaciones incrementales:** capacidad de actualizar de forma incremental destinos de datos en lugar de reconstruirlos cada vez desde cero.
- **Interfaz:** algunos productos ofrecen un entorno de desarrollo visual integrado que hace que sean más fáciles de usar. Estas herramientas gráficas permiten al usuario definir los procesos de calidad de datos modelados como flujos de trabajo usando una interfaz de "arrastrar y soltar".
- **Repositorio de metadatos:** repositorio que almacena los esquemas de datos e información, sobre el diseño de los procesos de calidad de datos. Esta información se consume durante la ejecución de los procesos de calidad de datos.
- **Técnicas de rendimiento:** conjunto de características para acelerar los procesos de limpieza de datos y para garantizar la escalabilidad. Algunas técnicas importantes para mejorar el rendimiento son: la separación, el procesamiento en paralelo, los hilos, el *clustering* y el balanceo de carga.
- **Versiones:** mecanismo de control de versiones con opciones de control estándar (por ejemplo, el registro de entrada, salida) que permite a los desarrolladores realizar un seguimiento de las diferentes versiones de desarrollo del código fuente.
- **Biblioteca de funciones:** conjunto de funciones predefinidas, tales como convertidores de tipos de datos y funciones de normalización, que abordan problemas específicos de calidad de datos. Una característica importante es la posibilidad de ampliar la biblioteca de funciones con nuevas funciones. Esto se puede lograr a través de un lenguaje de programación o mediante la adición de funciones externas de una librería de enlace dinámico (DLL, por sus siglas en inglés).
- **Lenguaje vinculante:** soporte de lenguaje de programación integrado para desarrollar nuevas funciones, con el fin de ampliar la biblioteca de funciones. Este soporte puede variar desde un lenguaje de programación existente (como *C* o *Perl*) a uno nativo.
- **Depuración y rastreo:** rastreo de gestiones que documentan la ejecución de los programas de calidad de datos con información útil (por ejemplo, tiempos de

inicio y fin de ejecución de rutinas importantes, el número de registros de entrada y salida).

- **Detección y manejo de excepciones:** las excepciones son el conjunto de registros de entrada para los cuales falla la ejecución por parte de los procesos de calidad de datos. El manejo de excepciones puede ser manual mediante una interfaz de usuario dada, o automático, al ignorar/borrar registros de excepción, o informar de ellos en un archivo de excepción o tabla.
- **Linaje de datos:** el linaje de datos o procedencia de datos, identifica el conjunto de elementos de datos fuente que produjo un elemento de datos dado (Cui, 2001).

Limpieza de datos

La existencia de problemas de calidad de datos, comúnmente llamados "datos sucios", degrada significativamente la calidad de la información, con un impacto directo en la eficiencia de la empresa que maneja dicha información (Barateiro y Galhardas, 2005). Una vez detectados los problemas de calidad en la fase de análisis de datos, entra en vigor el proceso de limpieza de datos con el objetivo de proporcionarle a los datos una mayor *"calidad"* para un buen desempeño posterior con los mismos. Este proceso forma parte de la fase de mejora de la metodología TDQM.

La limpieza de datos (también conocida como limpiador o fregador de datos) es el acto de detectar, eliminar y/o corregir *"datos sucios"* con el objetivo de obtener datos de alta calidad. La limpieza de datos no tiene solamente como objetivo limpiar datos, sino también dar consistencia a diferentes conjuntos de datos que hayan sido combinados a través de bases de datos independientes (Barateiro y Galhardas, 2005).

Este proceso de limpieza es una tarea crucial en diversos escenarios de aplicación. Dentro de una sola fuente de datos (por ejemplo, un listado de clientes), es importante su uso para corregir los problemas de integridad, estandarizar valores, rellenar los datos que faltan y consolidar ocurrencias duplicadas. La construcción de un almacén de datos (DW, acrónimo del inglés, *Data Warehouse*) (Chaudhuri y Dayal, 1997) requiere un importante paso denominado ETL (Extracción, Transformación y Carga), proceso que se encarga de extraer información de las fuentes de datos operacionales, transformar dicha información y posteriormente cargarla en el esquema de un almacén de datos. Varios procesos de migración de datos tienen como objetivo convertir los datos legados almacenados en fuentes con un cierto esquema, en las fuentes de datos destino cuyo esquema es distinto y predefinido (Carreira y Galhardas, 2004). Dentro de la metodología TDQM, la limpieza de datos adquiere una relevancia especial para la comunidad científica y de negocios; en este marco dicho proceso de limpieza es aquel

que trata de "precisar el grado de corrección en los datos y mejorar su calidad" (Fox *et al.*, 1994).

Dependiendo del contexto en el que se aplique la limpieza de datos, este proceso puede ser conocido bajo diferentes nombres; por ejemplo, cuando se detectan y eliminan registros duplicados dentro de un solo archivo, un registro de enlace (área estadística) o el proceso de eliminación de duplicados (área de base de datos) se llevan a cabo. En el contexto de almacenes de datos, los procesos ETL encierran tareas de limpieza de datos, y algunos autores como (Kimbal *et al.*, 1998) designan un almacenamiento específico para los datos, denominado área de organización o preparación de los datos (*data staging area*) en la arquitectura de almacenes de datos para la recopilación de los resultados intermedios de las transformaciones de limpieza de datos (Barateiro y Galhardas, 2005).

Algoritmos o métodos empleados en la limpieza de datos

Los métodos para llevar a cabo la limpieza de datos están estrechamente vinculados al área en que se aplica y el paso del proceso que se esté realizando. Sin embargo, se destacan algunos métodos generales muy usados:

Para la detección de errores *los métodos estadísticos* que aunque simples y rápidos pueden generar muchos falsos positivos, *los métodos de agrupamiento basados en distancias* cuya principal desventaja radica en su complejidad computacional, *los métodos basados en patrones y en reglas de asociación* que a partir del análisis de los registros que incumplen los patrones y las reglas descubiertas, detectan posibles errores (Marcus y Maletic, 2005).

También en Müller y Freytag (2003) se describe el método denominado *parsing* (Raman y Hellerstein, 2001), *las transformaciones a nivel de esquemas e instancias* (Sattler y Schallehn 2001), *el reforzamiento de las restricciones de integridad* (Suzanne *et al.*, 2001), *el método de las vecindades ordenadas* (Lee *et al.*, 1999, Hernández y Stolfo, 1998) y otros (Ananthakrishna *et al.*, 2002; Bilenko y Mooney, 2003; Lehti y Fankhauser, 2005; Monge y Elkan, 1997; Arasu *et al.*, 2009) que utilizan diferentes enfoques para realizar la eliminación de duplicados.

Herramientas comúnmente usadas para la limpieza de datos

Varias aplicaciones sofisticadas de software están disponibles para limpiar datos utilizando funciones específicas, normas y tablas de consulta. En el pasado, esta tarea se realizaba manualmente y por tanto se encontraba sujeta a errores humanos. En la actualidad se pueden utilizar herramientas que aplican técnicas de limpieza de datos

automatizadas de manera que facilitan este trabajo, las mismas se clasifican como herramientas comerciales y de investigación (Barateiro y Galhardas, 2005). Dentro de las herramientas comerciales se encuentran: *DataBlade, dfPower, ETLQ, ETI*DataCleanser, Firstlogic, NaDIS, QuickAddress Batch, Sagent, Trillium, WizRule, WizWhy, DataFusion, Hummingbird ETL, Informatica ETL, SQL Server 2000 DTS, SQL Server 2005, SQL Server 2008, Sunopsis, Centrus Merge/Purge, ChoiceMaker, DeDupe, DoubleTake, Identity Search Server, MatchMaker, Merge/Purge Plus, WizSame, DataStage, PureName PureAddress, Integrity, Oracle, Oracle 10g, SSA-Name/Data Clustering Engine, d. Centric, reUnion and MasterMerge, PureIntegrate, TwinFinder, Data Tools Twins, NoDupes, DeDuce, DataCleaner, Pentaho Data Integration* y *RapidMiner*. Dentro de las herramientas de investigación se destacan: *IntelliClean, Flamingo Project, TranScm, Potter´s Wheel, Clio, Ajax, Arktos* y *FraQL*, entre otras (Barateiro y Galhardas, 2005).

Tomando en cuenta las herramientas mencionadas anteriormente, a continuación se muestra un resumen de los principales rasgos con los que cuentan las herramientas de calidad de datos, analizando diferentes funcionalidades que poseen (ver Tabla 9). Varias de las herramientas no han vuelto a ser incluidas en la Tabla 9 debido a su anterior inclusión en la Tabla 8 porque tienen la capacidad de realizar ambos procesos (perfilado y limpieza de datos). Donde el valor Y significa soportado, X no soportado, - información desconocida, N nativo, DB bases de datos relacionales, FF archivos planos, G gráfica; NG no gráfica e Informix significa información mixta respectivamente.

Tabla 9: Funcionalidades generales de herramientas de calidad de datos comerciales y de investigación usadas en la limpieza de datos Basado (Barateiro and Galhardas, 2005)

Herramienta	Origen de datos	Capacidad de extracción	Capacidad de carga	Actualizaciones incrementales	Interfaz	Repositorio de metadatos	Técnicas de rendimiento	Versiones	Biblioteca de funciones	Lenguaje vinculante	Depuración	Manejo de errores	Linaje de datos
DataFusion	DB	Y	DB	Y	G	-	Y	Y	Y	N	Y	X	-
Centrus Merge/Purge	DB	-	-	-	G	-	-	-	-	-	-	-	-
DataBlade	Informix	-	Informix	-	G	-	-	-	-	-	Y	X	X
ChoiceMaker	DB,FF	-	-	-	G	Y	Y	-	Y	N	Y	Y	Y
ETI*Data Cleanser	Varias	-	-	-	G	Y	Y	-	Y	Y	-	Y	-
DeDupe	DB	-	-	-	G	-	-	-	-	-	-	-	-

NaDIS	-	X	-	X	G	X	-	-	X	X	-	X	X
QuickAddress Batch	ODBC	X	-	X	G	X	-	-	X	X	-	X	X
DoubleTake	ODBC	-	-	-	G	-	-	-	-	Y	-	-	-
Identity Search Server	DB	-	-	Y	G	Y	-	-	-	-	-	-	-
MatchMaker	DB	-	-	-	G	-	-	-	-	-	Y	-	-
Merge/Purge Plus	-	-	-	-	-	-	-	-	-	-	-	-	-
WizSame	DB,FF	-	-	-	G	-	Y	-	-	-	-	-	-
PureName PureAddress	-	-	-	-	-	-	-	-	-	-	-	-	-
Integrity	-	-	-	-	-	-	-	-	-	-	-	-	-
SSA-Name/Data Clustering Engine	-	-	-	-	-	-	-	-	-	-	-	-	-
d. Centric	-	-	-	-	-	-	-	-	-	-	-	-	-
PureIntegrate	-	-	-	-	-	-	-	-	-	-	-	-	-
TwinFinder	-	-	-	-	-	-	-	-	-	-	-	-	-
Data Tools Twins	-	-	-	-	-	-	-	-	-	-	-	-	-
NoDupes	-	-	-	-	-	-	-	-	-	-	-	-	-
DeDuce	-	-	-	-	-	-	-	-	-	-	-	-	-
IntelliClean	DB	-	DB	-	NG	-	-	-	-	-	Y	X	-
Flamingo Project	DB	-	DB	-	NG	-	-	-	-	-	-	-	-
TranScm	Y	-	-	-	G	-	-	-	Y	N	-	-	-
Clio	DB, XML	X	-	X	G	Y	Y	X	-	-	Y	-	-
Ajax	DB, FF	Y	DB	X	NG	X	Y	X	Y	Java	Y	Y	Y
Arktos	JDBC	-	JDBC	-	G	-	-	X	-	-	Y	Y	-
FraQL	-	-	-	-	NG	-	-	-	Y	N	-	-	-
Pentaho Data Integration	Varias	Y	Y	-	G	-	-	-	-	-	-	-	-
RapidMiner	-	-	-	-	G	-	-	-	-	-	-	-	-

Basado (Barateiro and Galhardas, 2005)

CONCLUSIONES

En el presente trabajo se expusieron las principales definiciones del término calidad de datos y su ciclo de vida a través de la caracterización de las cuatro fases que componen la metodología TDQM: definición, medición, análisis y mejora. Además, se presentaron las técnicas para identificar las dimensiones de calidad de datos en un contexto específico, haciendo énfasis en las dimensiones exactitud, completitud, consistencia y las relacionadas con el tiempo y se hizo corresponder este proceso con la fase de definición. Igualmente se expresaron las diversas formas de medir la calidad de los datos a través de la definición de métricas propuestas por varios autores, las cuales pertenecen a la fase de medición. También se expuso una taxonomía de errores comunes y las principales herramientas que se pueden usar para el perfilado de datos como parte de la fase de análisis. Finalmente, se describió el proceso de limpieza de datos el cual se ejecuta durante la fase de mejora de la metodología TDQM y se mencionaron los diferentes algoritmos, métodos y herramientas que generalmente se utilizan en esta etapa.

REFERENCIAS BIBLIOGRÁFICAS

Abelló, A., Darmont, J., Etcheverry, L., Golfarelli, M. *et al.*: *Fusion cubes: towards self-service business intelligence*. 2013.

Amón, I., Moreno, F. & Echeverri, J.: "Algoritmo fonético para detección de cadenas de texto duplicadas en el idioma español". *Revista Ingenierías Universidad de Medellín*, 11, pp. 127-138, 2012.

Ananthakrishna, R., Chaudhuri, S. & Ganti, V: "Eliminating fuzzy duplicates in data warehouses". Proceedings of the 28th VLDB Conference, Hong Kong, China, 2002.

Arasu, A., Chaudhuri, S. & Kaushik, R: *Learning string trasnformations from examples.* VLDB' 09, *august* 24-28, Lyon, France, 2009.

Arenas, M., Bertossi, L. & Chomicki, J.: "Consistent query answers in inconsistent databases. Proceedings of the eighteenth". ACM SIGMOD-SIGACT-SIGART, Symposium on principles of database systems. ACM, pp. 68-79, 1999.

Atzeni, P. & De Antonellis, V.: *Relational database theory*, Benjamin-Cummings Publishing Co., Inc, 1993.

Ballou, D., Wang, R., Pazer, H. & Tayi, G. K.: "Modeling information manufacturing systems to determine information product quality". *Management Science*, 44, pp. 462-484, 1998.

Ballou, D. P. & Pazer, H. L.: "Modeling data and process quality in multi-input, multi-output information systems". *Management Science,* 31, pp.150-162, 1985.

Barateiro, J. & Galhardas, H.: A Survey of Data Quality Tools. *Datenbank-Spektrum*, 14, 48, 2005.

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A.: "Methodologies for data quality assessment and improvement". *ACM Computing Surveys (CSUR)*, 41, 16.}, 2009.

Bilenko, M. & Mooney, R. J.: "On evaluation and training-set construction for duplicate detection". *Proc. of the KDD-2003 Workshop on Data Cleaning, Record Linkage, and Object Consolidation*, pp.7-12, 2003.

Bobrowski, M., Marré, M. & Yankelevich, D.: "Measuring data quality". Universidad de Buenos Aires, Argentina, 1999.

Bovee, M., Srivastava, R. & Mak, B.: "A conceptual framework and belief-function approach to assessing overall information quality". Proceedings of the 6th International Conference on Information Quality, 2001.

Calı, A., CALVANESE, D., DE GIACOMO, G. & LENZERINI, M.: "Data integration under integrity constraints". *Information Systems,* 29, pp.147-163, 2004.

Carreira, P. & Galhardas, H. “Efficient development of data migration transformations”. Proceedings of the 2004 ACM SIGMOD international conference on Management of data, 2004.

Codd, E. F.: “Relational database: a practical foundation for productivity”. *Communications of the ACM,* 25, pp. 109-117, 1982.

Cohen, W. W., Ravikumar, P. & Fienberg, S. E. *A Comparison of String Distance Metrics for Name-Matching Tasks.* Proceedings of II Web, 2003.

Cui, Y.: *Lineage tracing in data warehouses.* Stanford University, 2001.

Chapman, A. D.: *Principles of data quality*. GBIF, 2005.

Chaudhuri, S. & Dayal, U.: “An overview of data warehousing and OLAP technology”. *ACM Sigmod record,* 26, pp. 65-74, 1997.

Chaudhuri, S., Ganti, V. & Xin, D.: *Exploiting web search to generate synonyms for entities,* Madrid, Spain, ACM, 2009.

Chen, C. C. & Tseng, Y.-D.: “Quality evaluation of product reviews using an information quality framework”. *Decision Support Systems*, 50, pp. 755-768, 2011.

Dani, M. N., Faruquie, T. A., Garg, R., Kothari, G. *et al*: “A knowledge acquisition method for improving data quality in services engagements”. Services Computing (SCC), 2010 IEEE International Conference on, IEEE, pp. 346-353, 2010.

Deming, W. E.: Out of the crisis, Massachusetts Institute of Technology. *Center for advanced engineering study*, Cambridge, 1986.

Doggett, A. M. “Root cause analysis: A framework for tool selection”. *Quality Management Journal,* 12, 34, 2005.

Dunn, H. L.; “Record linkage”. *American Journal of Public Health and the Nations Health*, 36, pp. 1412-1416, 1946.

Else, W. I.: *The Complete Soundex Guide: Discovering the Rules Used by the Census Bureau and the Immigration and Naturalization Service when Those Organizations Indexed Federal Records; an Unofficial Supplement to National Archives' Federal Population Census and Immigrant and Passenger Arrivals Publications and More*. Closson Press, 2002.

Emran, N. A., Abdullah, N. & Mustafa, N.: *A Review of Failure Handling Mechanisms for Data Quality Measures.* 2012.

Eppler, M. J. & Wittig, D.: “Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years”. *IQ*, pp. 83-96, 2000.

Fellegi, I. P. & Holt, D.: “A systematic approach to automatic edit and imputation”. *Journal of the American Statistical association,* 71, pp. 17-35, 1976.

Fisher, C. W., Chengalur-Smith, I. & Ballou, D. P. 2003. "The impact of experience and time on the use of data quality information in decision making". *Information Systems Research,* 14, 170-188.

Fisher, C. W., Lauria, E. J. & Matheus, C. C: "An accuracy metric: Percentages, randomness, and probabilities". *Journal of Data and Information Quality (JDIQ),* 1, 16, 2009.

Fox, C., Levitin, A. & Redman, T.: "The notion of data and its quality dimensions". *Information processing & management,* 30, pp. 9-19, 1994.

Francalanci, C. & Pernici, B.: "Data quality assessment from the user's perspective". Proceedings of the 2004 international workshop on Information quality in information systems, ACM, pp.68-73, 2004.

Gálvez, C.: "Identificación de nombres personales por medios de sistemas de codificación fonética". *Encontros Bibli,* Segundo Semestre, pp. 105-116, 2006.

Gelbukh, A., Jimenez, S., Becerra, C. & González, F.: "Generalized Monge-Elkan method for approximate text string comparision". 10th International Conference on Computational Linguisticsand Intelligent Text Processing, pp. 559-570, 2009.

Gill, L.: *OX-LINK: the Oxford medical record linkage system*. Citeseer, 1997.

González, Y. I.: "Descubrimiento de llaves foráneas en colecciones de datos: reglas de inclusión". *Serie Científica,* 3, 2010.

Hall, P. & Dowling, G.: "Approximate string matching". *ACM Computing Surveys,* 12, PP. 381-402, 1980.

Hernández, M. A. & STOLFO, S. J.: "Real world Data is Dirty: Data Cleansing and The Merge-Purge Problem". *Journal of data mining and Knowlwgde Discovery,* 2, 1998.

Huang, K.-T., Lee, Y. W. & Wang, R. Y.: *Quality information and knowledge*, Prentice Hall PTR, 1998.

Iso, I. O. F. S. & Ansi, A. N. S. I.: *ISO International Standard: Database Language SQL – part 2*, 1999.

Jarke, M., Lenzerini, M., Vassiliou, Y. & Vassiliadis, P.: *Fundamentals of Data Warehouses*, Springer Verlag, 1995.

Jebamalar-Tamilselvi, J. & Saravanan, V.: "A Unified Framework and Sequential Data Cleaning Approach for a Data Warehouse". *International Journal of Computer Science and Network Security,* 8, 2008.

Juran, J. M. & Gryna, F. M.: *Quality planning and analysis*, McGraw-Hill, 1980.

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K. & Lee, D.: "A taxonomy of dirty data". *Data mining and knowledge discovery,* 7, PP. 81-99, 2003.

Kimbal, R., Reeves, L., Ross, M. & Thornthwaite, W.: *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses.* John Wiley & Sons, 1998.

Koronios, A., Lin, S. & Gao, J.: "A data quality model for asset management in engineering organisations". *IQ,* 2005.

Kovac, R., Lee, Y. W. & Pipino, L.: "Total Data Quality Management: The Case of IRI". *IQ*, pp. 63-79, 1997.

Lee, M. L., Lu, H., Ling, T. W. & Ko, Y. T.: Cleansing data for mining and datawarehousing. 1999.

Lee, Y. W., Pipino, L. L., Funk, J. D. & Wang, R. Y.: "Journey to Data Quality". *The MIT Press,* 36, 2006.

Lehti, P. & Fankhauser, P.: "A Precise Blocking Method for Record Linkage". *Lecture Notes Computer Science,* 3589, pp. 210-220, 2005.

Levenshtein, V. I.: "Binary Code capable of correcting deletions, insertions and reversal". *Soviet Physics Doklady,* 163, pp. 845-848, 1965.

Li, C., Wang, B. & Yang, X. Vgram: Improving Performance of Approximate Queries on String Collections Using Variable-Length Grams. *In:* ACM, ed. VLDB'07, September 23-28, Vienna, Austria, 2007.

Liu, L. & Chi, L.: Evolutionary data quality. Proceedings of the 7th international conference on information quality (IQ), MIT, Cambridge, USA, 2002.

Madnick, S. & Wang, R.: Introduction to total data quality management (TDQM) research program. TDQM-92-01, Total Data Quality Management Program, MIT Sloan School of Management, 1992.

Madnick, S. E., Wang, R. Y., Lee, Y. W. & Zhu, H.: "Overview and framework for data and information quality research". *Journal of Data and Information Quality (JDIQ),* 1, 2, 2009.

Marcus, A. & Maletic, J. I.: *Cap. 2 A Prelude to Knowledge Discovery. The Data Mining and Knowledge Discovery Handbook*, Springer, 2005.

Marshall, B.: Data quality and data profiling: a glossary of terms. 2007.

Maydanchik, A.: *Data quality assessment,* Technics publications. 2007.

Moges, H.-T., Dejaeger, K., Lemahieu, W. & Baesens, B.: "A multidimensional analysis of data quality for credit risk management: New insights and challenges". *Information & Management,* 50, pp. 43-58, 2007.

Monge, A. E. & Elkan, C. P.: "The field matching problem: Algorithms and applications". *In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 267—270, 1996.

_____: “An efficient domain-independent algorithm for detecting approximately duplicate database tuples”. *Proceedings of the SIGMOD 1997 workshop on data mining and knowledge discovery*, 1997.

Moore, D. S. & Mccabe, G. P.: *Introduction to the Practice of Statistics*, WH Freeman/Times Books/Henry Holt & Co, 1989.

Müller, H. & Freytag, J.-C.: *Problems, methods, and challenges in comprehensive data cleansing*, Professoren des Inst. Für Informatik, 2005.

Müller, H. & Freytag, J. C.: “Problems, Methods, and Challenges in Comprehensive Data Cleansing”. *Technical Report HUB-IB-164, Humboldt University Berlin, 2003.*

Naumann, F.: *Quality-driven query answering for integrated information systems*, Springer Science & Business Media, 2002.

Newcombe, H. B. & Kennedy, J. M.: “Record linkage: making maximum use of the discriminating power of identifying information”. *Communications of the ACM,* 5, pp. 563-566, 1962.

Oliveira, P., Rodrigues, F., Henriques, P. & Galhardas, H.: “A taxonomy of data quality problems”. 2nd Int. Workshop on Data and Information Quality, pp. 219-233, 2005.

Olson, J. E.: *Data quality: the accuracy dimension*. Morgan Kaufmann, 2003.

Pandey, R. K.: Data Quality in Data warehouse: problems and solution. 2014.

Patman, F. & Shaefer, L.: Is Soundex Good Enough for You? On the Hidden Risks of Soundex-Based Name Searching© 2001-2003 Language Analysis Systems. Inc, 2003.

Philips, L.: “Hanging on the metaphone”. *Computer Language,* 7, 1990.

_____: “The double metaphone search algorithm”. *C/C++ users journal,* 18, pp. 38-43, 2000.

Pipino, L. L., Lee, Y. W. & Wang, R. Y.: “Data quality assessment”. *Communications of the ACM,* 45, pp. 211-218, 2002.

Porrero, B. L.: *Limpieza de datos: Remplazo de valores ausentes y estandarización*. 2011.

Price, R. & Shanks, G.: “The impact of data quality tags on decision-making outcomes and process”. *Journal of the Association for Information Systems,* 12, 1, 2011.

Raghavan, H. & Allan, J.: “Using soundex codes for indexing names in ASR documents”. Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004, Association for Computational Linguistics, pp. 22-27, 2004.

Rahm, E. & Do, H. H.: “Data cleaning: Problems and current approaches”. *IEEE Data Eng. Bull.*, 23, pp. 3-13, 2000.

Raman, V. & Hellerstein, J. M.: “Potter's Wheel: An Interactive Data Cleaning System”. *The VLDB Journal*, pp. 381-390, 2001.

Redman, T. C.: *Data quality for the information age*. Boston: Norwood: Artech House, 1996.

Rema, A., Vijil, C., Prasad, M. D. & Raghuram, K.: *Rule based synonyms for entity extraction from noisy text*, Singapore, ACM, 2008.

Sadiq, S., Yeganeh, N. K. & Indulska, M.: “20 years of data quality research: themes, trends and synergies”. Proceedings of the Twenty-Second Australasian Database Conference-Volume 115. Australian Computer Society, Inc., pp. 153-162, 2011.

Salomone, S., Hyland, P. & Murphy, G. D.: Perceptions of data quality dimensions and data roles. 25th ANZAM Conference, 2011.

Sattler, K. U. & Schallehn, E.: “A Data Preparation Framework based on a Multidatabase Language”. *International Database Engineering Applications Symposium (IDEAS)*, 2001.

Scannapieco, M. & Catarci, T.: “Data quality under a computer science perspective”. *Archivi & Computer*, 2, pp. 1-15, 2002.

Scherbina, A.: “Clustering of Web Access Sessions”. *Lecture Notes in Computer Science*, 2005.

Sessions, V. & Valtorta, M.: “Towards a method for data accuracy assessment utilizing a Bayesian network learning algorithm”. *Journal of Data and Information Quality (JDIQ)*, 1, 14, 2009.

Shankaranarayanan, G. & Cai, Y.: “Supporting data quality management in decision-making”. *Decision Support Systems*, 42, pp. 302-317, 2006.

Shankaranarayanan, G. & Zhu, B.: “Data quality metadata and decision making. System Science (HICSS)”, 45th Hawaii International Conference on, IEEE, pp. 1434-1443, 2012.

Smith, T. F. & Waterman, M. S.: “Identification of common molecular subsequences”. *Journal Molecular Biology*, pp. 195-197, 1981.

Stanier, A.: “How accurate is Soundex matching”. *Computers in Genealogy*, 3, pp. 286-288, 1990.

Strong, D. M., Lee, Y. W. & Wang, R. Y.: “Data quality in context”. *Communications of the ACM*, 40, pp.103-110, 1997.

Suzanne, M. E., Brandt, S. M., Robinson, J. S., Sutherland, I., et al: “Adapting integrity enforcement techniques for data reconciliation”. *Inf. Syst.*, 26, pp. 657-689, 2005.

Taft, R. L.: *Name search techniques*, Bureau of Systems Development, New York State Identification and Intelligence System, 1970.

Tayi, G. K. & Ballou, D. P.: "Examining data quality". *Communications of the ACM*, 41, pp. 54-57, 1998.

Türker, C. & Gertz, M.: "Semantic integrity support in SQL: 1999 and commercial (object-) relational database management systems". *The VLDB Journal*, 10, pp. 241-269, 2001.

Ukkonen: "Approximate string matching with q-grams and maximal matches". *Theoretical Computer Science*, pp. 191-211, 1992.

Wand, Y. & Wang, R. Y.: "Anchoring data quality dimensions in ontological foundations". *Communications of the ACM*, 39, pp. 86-95, 1996.

Wang, R. & Lee, Y.: *Integrity Analyzer: A Software Tool for Total Data Quality Management.* Cambridge Research Group, Cambridge, MA, 1998.

Wang, R. Y.: "A product perspective on total data quality management". *Communications of the ACM*, 41, pp. 58-65, 1998.

Wang, R. Y., Storey, V. C. & Firth, C. P.: "A framework for analysis of data quality research". *Knowledge and Data Engineering, IEEE Transactions on*, 7, pp. 623-640, 1995.

Wang, R. Y. & Strong, D. M.: "Beyond accuracy: What data quality means to data consumers". *Journal of Management Information Systems*, pp. 5-33, 1996.

Watts, S., Shankaranarayanan, G. & Even, A.: "Data quality assessment in context: A cognitive perspective". *Decision Support Systems*, 48, pp. 202-211, 2009.

Watts, S. & Zhang, W.: *Knowledge adoption in online communities of practice*, System d'Information et Management, 2004.

Wijnhoven, F., Boelens, R., Middel, R. & Louissen, K.: Total Data Quality Management: A Study of Bridging Rigor and Relevance, 2007.

Wingkvist, A., Ericsson, M., Lincke, R. & Lowe, W.: "A metrics-based approach to technical documentation quality. Quality of Information and Communications Technology (QUATIC)". Seventh International Conference on the, IEEE, pp. 476-481, 2010.

Winkler, W. E.: "Overview of record linkage and current research directions". *Technical Report RR2006/02*, 2006.

Winkler, W. E. & Census, U. S. B. O. T.: Improved decision rules in the Fellegi-Sunter model of record linkage, 1993.

Yeganeh, N. K., Sadiq, S. & Sharaf, M. A.: “A framework for data quality aware query systems”. *Information Systems*, 2014.

Zahedi, F.: *Quality information systems*, Boyd & Fraser Publishing Co, 1995.

Zeeh, C. The lempel ziv algorithm. URL: http://w3studi. informatik. uni-stuttgart. de/~ zeehca/Seminar/LempelZivReport. pdf [accessed November 3, 2003], 2003.

Zhu, X. & Gauch, S.: “Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web”. Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp. 288-295, 2000.

Zobel, J. & Dart, P. Phonetic String Matching: Lessons from Information Retrieval. Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval, 1996.