UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS
VERITATE SOLA NOBIS IMPONETUR VIRILIS TOGA. 1948

Facultad de Matemática, Física y Computación
Centro de Estudios de Informática
Departamento de Ciencia de la Computación

TRABAJO DE DIPLOMA

"Estudio de las metodologías para la determinación de la calidad de datos en volúmenes de datos"

Autor:

Mario Enrique Landin Alvarez

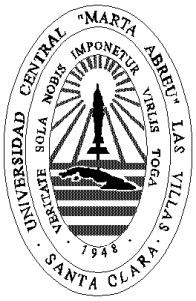
Tutora:

Dra. Beatriz López Porrero

Santa Clara

2015

"Año 57 de la Revolución"



El que subscribe: Mario Enrique Landin Alvarez, hago constar que el presente trabajo de diploma titulado: **Estudio de las metodologías para la determinación de la calidad de datos en volúmenes de datos**, fue realizado en la Universidad Central "Marta Abreu" de Las Villas como parte de la culminación de estudios de la especialidad de Licenciatura en Ciencia de la Computación, autorizando a que el mismo sea utilizado por la Institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos, ni publicados sin autorización de la Universidad.

Firma del Autor

Los abajo firmantes certificamos que el presente trabajo ha sido realizado según acuerdo de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

Firma del Tutor

Firma del Jefe de Departamento donde se defiende el trabajo

Firma del Responsable de Información Científico-Técnica

"En ambientes complejos, el seguir un plan al pie de la letra produce el producto que pretendíamos pero no el producto que necesitamos"

Highsmith

AGRADECIMIENTOS

En primer lugar quisiera agradecer a mi tutora, la Dra. Beatriz López Porrero y a el profesor Dr. Daniel Gálvez Lio, por su amistad, ayuda y confianza que han depositado en mí durante todos estos cinco años.

A mis padres y mis dos hermanas por su apoyo y consejos siempre acertados, tanto en lo académico como en lo personal.

A mi tía Iliana Landin, quien quiso verme temblando un poco hoy, a mis abuelos, a Liliana y al resto de la familia que de una forma u otra siempre estuvieron al tanto de mis estudios.

Agradecer a todos mis compañeros y amigos que me han acompañado durante estos años, en especial a los aliados en el Dota II: los hermanos Pabel y Pablo Ulacia, Pedro Alejandro y Omar. A los siempre amigos: Yordan, Ernesto Julio y Lisvandy.

Agradecer a Jose Carlos por su paciencia para conmigo.

Deseo manifestar mi más sincero agradecimiento a cuantas personas han hecho posible con su aportación, de forma directa o indirecta en la realización del presente trabajo.

A todos muchas gracias.

RESUMEN

El tema está enfocado a describir y analizar los principales elementos de las investigaciones sobre calidad en bases de datos, fundamentalmente desde la perspectiva de los negocios, relacionando aspectos tales como las dimensiones, las métricas y las metodologías.

La calidad de datos se refiere a los procesos y técnicas enfocados a medir y mejorar la calidad de los datos existentes en volúmenes de datos. Para que un proceso de calidad de datos sea realmente eficaz, este deberá 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 establecimiento del perfil de los datos, la estandarización o normalización, la correspondencia y consolidación; pasos que generarán reportes para dar seguimiento a los progresos y permitir la mejora continua de su calidad.

En esta investigación se estudia la calidad de datos con el objetivo de medir o establecer el nivel de calidad de los datos en grandes volúmenes de datos.

ABSTRACT

The subject is aimed to describing and analyzing the main elements of research about data quality in databases mainly from business's perspective, relating aspects such as dimensions, metrics and methodologies.

Data Quality refers to the processes and techniques aimed to measuring and improving the efficiency of existing data in the database. For a data quality process is really effective it must be repeatable and easy to understand, so that will generate a cycle of improvement and each time that is executed data they are obtained with higher quality. The process should include the establishment of the profile data, standardization or normalization, correspondence and consolidation; steps that will generate reports to track progress and enable continuous improvement of its quality.

In this research, the quality of data is studied in order to measure or determine the level of quality of data in large volumes of data.

TABLA DE CONTENIDOS

INTRODUCCIÓN

Es una realidad que parte de los datos almacenados por las organizaciones, contienen errores y estos pueden conducir a tomar decisiones erróneas, ocasionando pérdidas de tiempo, dinero y credibilidad (URIBE, 2010), de ahí que la calidad de los datos ha sido una preocupación desde el inicio de la automatización del procesamiento de los datos. Los usuarios de los sistemas de información aprendieron de forma rápida las fatales consecuencias que traen los datos erróneos en las bases de datos.

En la sociedad actual, la exploración de los datos para extraer información o conocimiento para el soporte a la toma de decisiones es sinónimo de éxito, sin embargo, son varios los problemas que pueden afectar la calidad de los datos, provocando un impacto negativo en las decisiones tomadas. En un contexto ideal, los datos almacenados no deberían contener errores, pero es innegable que los errores son una realidad y merecen toda la atención. Los datos pueden presentar problemas de diferentes tipos como duplicados, valores faltantes, valores atípicos y algunos de ellos incorrectos o variaciones tipográficas (URIBE, 2010).

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

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

Dado lo antes expuesto se plantea el siguiente problema de investigación: ***Estudio de las metodologías para la determinación de la calidad de datos en volúmenes de datos***.

Para dar solución al problema de investigación se plantea como **objetivo general**: *Realizar un estudio de las metodologías para el esclarecimiento del proceso de determinación de la calidad de datos en volúmenes de datos.*

Para dar cumplimiento al objetivo propuesto se proponen los siguientes **objetivos específicos:**

- *Definir los términos relacionados con el proceso de calidad de datos, desde la perspectiva de cómo medirla en volúmenes de datos.*
- *Describir las principales metodologías para la determinación de la calidad de datos.*
- *Seleccionar y aplicar una metodología a la base de datos del SIGENU.*

Para dar solución al problema de investigación y a los objetivos específicos, se le darán respuestas a las siguientes **preguntas de investigación:**

- *¿Cuáles son las principales áreas en que es aplicable la calidad de datos?*
- *¿Cuáles son las principales dimensiones y métricas a tratar en la determinación de la calidad de un conjunto de datos?*
- *¿Cuáles son las principales metodologías para medir la calidad de datos en grandes volúmenes de datos?*

Justificación de la investigación:

En el laboratorio de Base de datos existe un antecedente en el tema de limpieza de datos, en que se trabajó en la perspectiva de las bases de datos, expresada en la determinación de la calidad en los valores e instancias, del cual se derivaron herramientas, entre ellas una para el análisis de datos (*data profiling*), que indiscutiblemente representa una gran ayuda para las investigaciones en el tema de calidad de los datos (Porrero, 2011). Además es necesario contar con datos coherentes, inequívocos, completos y precisos que sean confiables y apoyen las decisiones estratégicas de los sistemas empresariales. El tema ofrecerá resultados dirigidos al personal que trabaja en el sector empresarial, donde los datos seguros juegan un papel fundamental en la toma de decisiones y al personal que se dedica a la producción de software, específicamente a la ingeniería de requisitos en la construcción de aplicaciones. Contribuirá significativamente a un uso más eficiente y efectivo de las tecnologías de la información, lo cual constituye impactos económicos y sociales.

Organización del informe

La tesis está estructurada en tres capítulos.

En el capítulo uno: Fundamentación Teórica, se realiza un estudio de la apreciación global del concepto de calidad de datos y de los beneficios de realizarla. Se expondrá una taxonomía de los conceptos relacionados con la calidad de datos, teniendo en cuenta para ello las distintas metodologías, métricas y dimensiones existentes en la literatura. Aborda también las áreas de investigación en las cuales se aplica la calidad de datos. El capítulo termina con la definición y las características de los conceptos de dimensiones y métricas antes mencionados.

En el capítulo dos: Análisis de las Metodologías de Calidad de Datos, se realiza un análisis descriptivo de las distintas metodologías de calidad de datos, mostrando sus formas de trabajo, marcos o fases. Además de mostrar los tipos de datos asociados a dichas metodologías y las dimensiones que por lo general se les son considerados. Se efectúa una comparación entre algunas de las metodologías, en la cual se toman aspectos de gran importancia.

El capítulo tres: Aplicación de la Metodología TDQM al SIGENU, se aplica la metodología TDQM a un conjunto de datos. Se brinda una descripción del sistema que contiene o trabaja sobre la base de datos en cuestión. La metodología es aplicada a una de las bases de datos del sistema: la base de datos operacional y a uno de sus módulos de trabajo: cliente secretaria. El capítulo muestra además las entrevistas realizadas al caso de uso secretaria de las distintas facultades de la universidad. Se ofrece un conjunto de dimensiones de calidad de datos necesarios para medir dicha calidad en el conjunto de datos presentado y se definen las métricas necesarias y asociadas a esas dimensiones.

El documento culmina con las conclusiones y recomendaciones del autor.

CAPÍTULO 1. FUNDAMENTACIÓN TEÓRICA

En este capítulo se aborda lo referente a las bases teóricas que fundamentan esta investigación. Se realiza una revisión en la literatura para determinar los conceptos relacionados con la calidad de los datos y los tipos de datos usados generalmente. Se mostrarán algunas de las áreas de investigación relacionadas con la calidad de datos y se expondrán los conceptos y características de las dimensiones y las métricas en el esclarecimiento del proceso de determinación de la calidad de datos.

1.1 Calidad de datos

Los datos representan objetos del mundo real. Dichas representaciones resultan ser aplicables en contextos de diferentes y variadas características. Por otro lado, los datos pueden ser almacenados o sometidos a algún proceso o transformación, siendo siempre de suma importancia para garantizar la sobrevivencia y éxito de las organizaciones (Gallo and Corrado, 2009).

Los primeros investigadores en tratar el problema de la calidad de datos fueron los estadísticos, al proponer una teoría matemática acerca de los duplicados en los datos estadísticos. Ellos fueron seguidos por investigadores del área de la gestión y la informática que a partir de 1990 definen el problema de la calidad de los datos electrónicos almacenados en bases de datos, mediante la definición, medición y mejora de dicha calidad.

El concepto de calidad de datos es muy amplio y está sujeto a diferentes definiciones e interpretaciones aunque todas convergen en que el concepto es relativo al uso del dato (Burns et al., 2000, Domingo et al., 2007, Redman, 2001, Tayi and Ballou, 1998). La calidad de datos ha sido esencialmente estudiada en las comunidades de investigadores de las bases de datos y de los negocios.

1.1.1 Definición de Calidad de datos

Se puede decir que no existe una definición simple para el concepto de “calidad de datos” (Batini and Scannapieca, 2006, Sedó, 2007).

La calidad de los datos depende de los procesos de diseño y producción involucrados en la generación de los datos. Entender el significado de calidad y la forma en que debe medirse es fundamental para lograr diseñar con calidad (Lidiansa, 2014).

En (Batini and Scannapieca, 2006), se aborda la definición planteando que es muy común hacer coincidir la calidad de los datos con su precisión, aclarándose que este concepto debe ser considerado multifacéticamente incluyendo además de la precisión otras dimensiones como la completitud, la consistencia y la actualidad.

En (Sedó, 2007), se ofrecen otras definiciones de calidad de datos, las cuales se muestran a continuación.

- La calidad de datos se refiere al grado de excelencia que posean los datos en relación con el ámbito en que se encuentren definidos.
- Calidad de datos es la totalidad de características que les permiten a los datos cumplir con un propósito específico.
- La calidad de datos es el grado de completitud, consistencia, persistencia y exactitud que hace un dato apropiado para un uso específico.
- Los datos pueden considerarse de calidad si cumplen con su propósito desde una perspectiva operativa, de toma de decisiones y planeación.

En (Wang and Strong, 1996), se propone un marco preliminar para la calidad de datos basados en las experiencias con los consumidores de datos, que incluye los aspectos siguientes:

- Los datos deben ser accesibles para los consumidores.
- Los consumidores deben ser capaces de interpretar los datos.
- Los datos deben ser relevantes para el consumidor.
- Los consumidores deben encontrar los datos exactos.

En el área de la calidad de datos se incluyen los aspectos que se muestran en la Tabla 1 (Gallo and Corrado, 2009).

Aspecto	Descripción
Dimensiones	Las mediciones sobre el nivel de calidad de los datos se aplican a las dimensiones de interés.
Metodologías	Proveen guías de acción.
Modelos	Representan las dimensiones y otros aspectos de la calidad de datos.
Técnicas	Proporcionan soluciones a problemas de calidad de datos.
Herramientas	Necesarias para que las metodologías y técnicas puedan llevarse a cabo de manera efectiva.

Tabla 1. Aspectos dentro de la calidad de datos.

1.2 Calidad de datos y tipos de datos

El objetivo final de una metodología de calidad es el análisis de los datos, que en general, describe objetos del mundo real en un formato que se puedan almacenar, recuperar y procesar por un software.

Los datos se distribuyen, cada vez más, en recursos heterogéneos y están representados en diferentes formatos que van desde los tipos de datos, no estructurados, por ejemplo: los sistemas de archivos, repositorios de documentos y la web, hasta tipos de datos estructurados, un ejemplo de estos son: los sistemas de gestión de bases de datos. En la literatura, las fuentes de datos son clasificadas según el nivel de la estructura que los caracteriza (Carlo et al., 2011a).

En el campo de la calidad de los datos, se distinguen tres tipos de datos: (Batini et al., 2009, Batini and Scannapieca, 2006, Carlo et al., 2011b, Jaaskelainen et al., 2011).

Datos estructurados

Los datos estructurados son agregaciones o generalizaciones de los elementos descritos por atributos definidos dentro de un dominio, estos dominios representan la gama de valores que pueden ser asignados a los atributos y por lo general corresponden a los tipos de datos de los lenguajes de programación, como son: los valores numéricos y las cadenas de texto. Las tablas relacionales y los datos estadísticos constituyen los tipos más comunes de estructuras que contienen datos estructurados.

Datos no estructurados

Los datos no estructurados son una secuencia genérica de símbolos, típicamente codificados en lenguaje natural, no están asociados a una estructura o tipo de dominio alguno. Algunos ejemplos típicos de datos no estructurados son los cuestionarios con textos libres que responden a preguntas abiertas o el cuerpo de un correo electrónico.

Datos semiestructurados

Son datos que tienen una estructura con cierto grado de flexibilidad. El lenguaje de marcas de uso común para representar este tipo de datos es XML.

En (Carlo et al., 2011a), se explica que las diferencias en el formato de los datos se reflejan necesariamente en los métodos y técnicas que las organizaciones utilizan con el fin de evaluar y mejorar la calidad de sus recursos de información.

Las técnicas, modelos, dimensiones y metodologías para la determinación de la calidad están estrechamente relacionadas al tipo de datos y se vuelven mucho más complejas para el tratamiento de datos semiestructurados y no estructurados. En (Crippa, 2006), se expresa que la mayoría de las contribuciones a la investigación sobre la calidad de los datos se centran en estos dos tipos de datos.

1.3 Problemas y consecuencias de la mala calidad en los datos

Los avances en la tecnología han permitido crear, almacenar y procesar grandes cantidades de datos. Conforme se ha incrementado el uso de la información almacenada, es evidente que los problemas relacionados con los datos inconsistentes afectan negativamente en términos de eficiencia y eficacia (Sedó, 2007).

Por lo general el tema de la calidad de los datos se ignora hasta que se convierte en un riesgo palpable o un obstáculo. La mala calidad de los datos en los sistemas de información es un tema que impacta negativamente hoy en día, por lo que se ha convertido en uno de los problemas más críticos que enfrentan las organizaciones y empresas en la actualidad.

En (Redman, 1998), se explican alguno de los impactos negativos provocados por la mala calidad en los datos, expresados a continuación.

- Clientes insatisfechos: sus datos personales, sus pedidos y facturas son incorrectos.
- Empleados insatisfechos: cometen errores o no conocen cierta información, lo que los hace cometer a su vez más errores.
- Toma de decisiones erróneas: los datos usados por los gerentes también pueden tener errores y es sabido que las decisiones no van a ser mejores que los datos en los que están basadas.

Los datos incorrectos se pueden generar de muchas formas, por errores en la entrada de datos, información incorrecta cargada desde formularios, discrepancias entre diferentes sistemas o datos incorrectos obtenidos por fuentes externas. También es común encontrar datos correctos mezclados con información desactualizada, lo que dificulta distinguir los datos válidos de los no válidos (Sedó, 2007).

Algunos de los problemas referidos a la calidad de los datos se resumen en la Tabla 2 (Sedó, 2007), (Fisher and Marinos, 2003).

Problemas	Definición
Problemas de estandarización	Errores de ortografía en los nombres y el uso de abreviaturas.
Inconsistencia en los formatos	Son datos que no cumplen con un formato definido. Se presentan con mayor frecuencia en los campos tipo fecha.
Datos incorrectos	Datos que cumplen con el formato pero no son válidos. Se presentan con mayor frecuencia en los campos tipo fecha.
Datos estáticos	Información desactualizada.

Tabla 2. Problemas referidos a la calidad de los datos.

Se alcanza un nivel aceptable de calidad si los datos se encuentran de acuerdo con las especificaciones y sirven para lo que fueron definidos (Batini and Scannapieca, 2006), (Strong et al., 1997).

1.4 Áreas de aplicación en la calidad de datos

Varias áreas del conocimiento y la ciencia han desarrollado en las últimas décadas paradigmas, modelos y metodologías que han demostrado ser de gran importancia en el área de investigación de calidad de datos.

Lograr calidad en los datos es una tarea compleja y multidisciplinaria, debido a su importancia, naturaleza, variedad de tipos de datos y sistemas de información que pueden estar implicados.

En (Batini and Scannapieca, 2006), se exponen algunos de los dominios de aplicación en calidad de datos que por su importancia y relevancia han ido creciendo en los últimos años. Algunos de estas áreas de aplicación se explican a continuación:

***e-Government* (Administración electrónica)**

El objetivo principal de todos los proyectos de administración electrónica es la mejora de la relación entre el gobierno, las agencias y los ciudadanos; así como entre agencias y empresas a través del uso de la información y la comunicación tecnológica. Este ambicioso objetivo se articula de la siguiente forma:

- La completa automatización de los procesos administrativos del gobierno que prestan servicios a los ciudadanos y las empresas y que implican el intercambio de datos entre los organismos gubernamentales.
- La creación de una arquitectura que mediante la conexión de las distintas agencias, les permita cumplir con sus procesos administrativos sin ninguna carga adicional para los usuarios que se benefician de ellos.
- La creación de portales que simplifican el acceso a servicios por parte de usuarios autorizados.

Los proyectos pertenecientes a la administración electrónica enfrentan la dificultad, de que una información similar alrededor de un ciudadano o negocio es probable que aparezca en varias bases de datos. Ligado a esto se insertan los errores que normalmente están presentes en las bases de datos, debido a la naturaleza del flujo administrativo y que los datos de varios ciudadanos no se actualizan durante períodos largos de tiempo. Además, pueden ocurrir errores vinculados a los datos personales de los ciudadanos y algunos de estos no se corrigen ni se detectan, por otro lado, los datos proporcionados por distintas fuentes en alguno de los casos difieren en su formato.

Ciencias de la vida

Los datos existentes, específicamente los datos biológicos almacenados en bases de datos, se caracterizan por la diversidad de tipos, los volúmenes muy grandes, la calidad muy variable y la disponibilidad a través de fuentes muy diversas.

***World Wide Web* (Red Informática Mundial)**

Los sistemas de información basados en la web se caracterizan por la presentación de una gran cantidad de datos, cuya calidad puede ser muy heterogénea dado el caso de que todas las organizaciones e individuos pueden crear un sitio web y cargar todo tipo de información en él, sin tener en cuenta el control de su calidad.

Estos sistemas presentan dos aspectos adicionales en relación a la calidad de los datos que las diferencian de las fuentes tradicionales de información, el primero de ellos es que un sitio web es una fuente de continua evolución de la información que no está vinculado a un tiempo de liberación fija de información, el segundo es que el proceso de actualización se puede producir en diferentes fases y es posible hacer correcciones a los datos ya publicados, creando así una necesidad más para establecer controles de calidad.

1.5 Áreas relacionadas con la calidad de datos

Algunas de las áreas de investigación relacionadas con la calidad de datos se resumen a continuación, (Batini and Scannapieca, 2006).

La estadística

Incluye un conjunto de métodos que se utilizan para recopilar, analizar, presentar e interpretar datos. La estadística ha desarrollado en los últimos dos siglos un amplio espectro de métodos y modelos que permitan expresar predicciones y formular decisiones en todos los contextos. La metodología estadística como base de análisis de datos tiene que ver con dos tipos básicos de problemas:

- Resume, describe y explora los datos.
- Usa los datos de la muestra para inferir la naturaleza del proceso que produjo dichos datos.

Dado que los datos de baja calidad son una inexacta representación de la realidad, se han desarrollado una variedad de métodos estadísticos para medir y mejorar la calidad de los datos.

Representación del conocimiento

Es el estudio de cómo el conocimiento sobre un dominio de aplicación puede ser representado, y el tipo de razonamiento que se puede hacer. El conocimiento sobre un dominio de aplicación puede ser representado a través del código de programa o implícitamente como patrones de activación en una red neuronal. Alternativamente, el área de representación del conocimiento asume una explícita y declarativa representación, en términos de una base de conocimientos que consiste en fórmulas lógicas o reglas expresadas en un lenguaje de representación. Proporcionar una rica representación del dominio de aplicación y ser capaz de razonar sobre ello se ha convertido en una influencia importante de muchas técnicas para la mejora de la calidad de los datos.

Minería de datos

Es un proceso analítico diseñado para explorar lo general de grandes conjuntos de datos en busca de patrones consistentes y/o relaciones sistemáticas entre los atributos/variables. La minería de datos exploratoria se define como el proceso preliminar de descubrimiento de patrones en un conjunto de datos utilizando resúmenes estadísticos, visualización y otros medios. Las técnicas de minería de datos pueden ser utilizadas en un amplio espectro de actividades para la mejora de la calidad de los datos.

Sistemas de gestión de la información

Se definen como sistemas que proporcionan la información necesaria para gestionar una organización de forma eficaz. Puesto que los datos y el conocimiento se están convirtiendo en los recursos más utilizados en la actualidad; tanto en los procesos de negocios operacionales y de decisión, los sistemas de gestión de información son cada vez más importantes, con funcionalidades y servicios que permiten controlar y mejorar la calidad de las fuentes de datos.

Integración de datos

Tiene el objetivo de construir y presentar una visión unificada de los datos presentes en las fuentes de datos. La integración de datos es considerada como una de las actividades básicas cuya finalidad es mejorar la calidad de los datos. Es un área de investigación autónoma y bien fundamentada que proporciona resultados correctos a las solicitudes, sobre la base de una caracterización de la calidad de los datos en las fuentes y la identificación y solución de conflictos en valores que se refieren a los mismos objetos del mundo real.

1.6 Dimensiones de la calidad de datos

Como se enuncia en el epígrafe: 1.1, la calidad de datos es un concepto que depende de las dimensiones que la definen.

Cada dimensión refleja un aspecto distinto de la calidad de los datos. Las mismas pueden estar referidas a la extensión de los datos: su valor. Existen varias dimensiones que reflejan distintos aspectos de los datos, dado que estos representan todo tipo de características de la realidad, desde espaciales y temporales, hasta sociales (Gallo and Corrado, 2009).

Algunas de las dimensiones que más frecuentemente, definen la calidad de los datos se resumen a continuación: (Batini et al., 2009, Batini and Scannapieca, 2006, Brackstone, 1999, Dejaeger et al., 2010, Domingo et al., 2007, Gallo and Corrado, 2009, Sivogolovko, 2011).

Exactitud

La exactitud se define como la cercanía que existe entre un valor v del mundo real y su representación v’ en los sistemas de información. Es tratada como una asociación correcta y precisa entre los estados del sistema de información y los objetos del mundo real.

Por ejemplo, si una base de datos almacena cierta información de una empresa y en la misma se encuentra registrada que existen diez computadoras almacenadas; efectivamente o físicamente deben existir esas diez computadoras en la empresa.

Se distinguen dos factores de exactitud: exactitud semántica y la exactitud sintáctica (Valverde et al., 2009), (Gallo and Corrado, 2009).

La exactitud semántica se refiere a la cercanía que existe entre un valor *v* y un valor real *v'*. Esta dimensión se mide fundamentalmente con valores booleanos indicando si es un valor correcto o no, para lo cual es necesario conocer cuáles son los valores reales a considerar.

Los tipos de errores que se identifican en este caso son:

- Registro inexistente.
- Defecto mal registrado.
- Valores fuera de referencial.

La exactitud sintáctica se refiere a la cercanía entre un valor *v* y los elementos de un dominio *D*. Esto es, si *v* corresponde a algún valor válido de *D* sin importar si ese valor corresponde a uno del mundo real.

Los tipos de errores que se identifican en este caso son:

- Valor fuera de rango.
- Estandarización.

Completitud

La completitud se puede definir como la medida en que los datos son de suficiente alcance y profundidad. Se trata como la capacidad del sistema de información de representar todos los estados significativos de una realidad dada.

Los tipos de errores que se identifican en este caso pueden encontrarse dentro del factor densidad:

- Valor nulo.
- Clasificación de defectos.

Por ejemplo, un cliente de un banco necesitará conocer el saldo de sus cuentas bancarias. Si esta no existiese, se tendría un problema en la completitud del dato, además de la insatisfacción del cliente.

Existen dos factores para expresar la completitud: cobertura y densidad.

La cobertura se refiere a la porción de los datos de la realidad que se encuentran contenidos en el sistema de información. Al igual que para la exactitud semántica, la cobertura involucra una

comparación del sistema de información con el mundo real. Una vez más se requiere de una referencia. Debido a que suele ser difícil obtenerla, la alternativa es estimar el tamaño de dicha referencia.

La densidad se refiere a la cantidad de información contenida y la información faltante acerca de las entidades en el sistema de información.

Consistencia

Esta dimensión hace referencia al cumplimiento de las reglas semánticas que son definidas sobre los datos. La inconsistencia de los datos se presenta cuando existe más de un estado del sistema de información asociado al mismo objeto de la realidad. Una situación que podría ocasionar inconsistencias en los datos es la incorporación de datos externos o con otros formatos.

Es necesario entonces la definición de estándares y protocolos para los datos. No es lo mismo que el género de una persona se almacene como “F" o como “Femenino" o “Fem". Debe definirse una forma única de almacenamiento de los datos.

La consistencia puede definirse en diferentes contextos (Loshin, 2006):

- Entre un conjunto de valores de atributos y otro atributo establecido en el mismo registro.
- Entre un conjunto de valores de atributos y otro conjunto de atributos en diferentes registros.
- Entre un conjunto de valores de los atributos y el mismo conjunto de atributos en el mismo registro en diferentes puntos en el tiempo.
- La coherencia también puede tener en cuenta el concepto de "razonabilidad", en el que se impone un rango de aceptabilidad de los valores de un conjunto de atributos.

Restricciones de integridad.

Las restricciones de integridad definen propiedades que deben ser cumplidas por todas las instancias de un esquema relacional. Se distinguen tres tipos de restricciones de integridad (Gallo and Corrado, 2009):

- Restricciones de dominio: se refiere a la satisfacción de reglas sobre el contenido de los atributos de una relación.
- Restricciones intra-relacionales: se refiere a la satisfacción de reglas sobre uno o varios atributos de una relación.
- Restricciones inter-relacionales: se refiere a la satisfacción de reglas sobre atributos de distintas relaciones.

Temporalidad o dimensiones relacionadas con el tiempo

Los cambios y actualizaciones de los datos son un aspecto importante a tener en cuenta cuando se valora su calidad. En determinados contextos un dato no actualizado es de mala calidad y puede llegar a ocasionar problemas graves.

Algunas de las dimensiones relacionadas con el tiempo se resumen en la Tabla 3.

Dimensión	Definición
Actualidad	Trata sobre la actualización de los datos y su vigencia. Esta dimensión puede ser medida de acuerdo a la fecha de la última actualización.
Volatilidad	Se refiere a la frecuencia con que los datos cambian en el tiempo. Una medida para esta dimensión es la cantidad de tiempo que los datos permanecen siendo válidos.

Tabla 3. Dimensiones relacionadas con el tiempo.

En (Sedó, 2007), se menciona la existencia de un grupo de características o atributos de los datos, los cuales son necesarios para lograr una medición de conformidad, este grupo de atributos puede variar dependiendo de los requerimientos del sistema de información. Debido a esto se hace necesario agrupar los atributos que poseen relación entre sí en dimensiones, para poder identificar de una mejor manera los problemas de la calidad de los datos.

En (Abate et al., 1998), se explica las ventajas que ofrece agrupar los atributos en dimensiones.

- Las dimensiones son más fáciles de entender que los atributos.
- Al unir atributos interdependientes en dimensiones, los analistas de calidad pueden consolidar y organizar la información necesaria para la interpretación y comprensión de los datos de una mejor manera.

- Las dimensiones ayudan a los analistas de calidad a identificar problemas sistemáticos de calidad en las aplicaciones.

En (Wang and Strong, 1996), se han agrupado 118 atributos en 20 dimensiones, que a su vez de agruparon en cuatro categorías: intrínsecas, contextuales, representación y accesibilidad.

Algunas de estas dimensiones son abordadas en (Caro et al., 2013), (Jaaskelainen et al., 2011)y se resumen en la Tabla 4.

Dimensión	Descripción
Credibilidad	Es el grado en el cual el dato tiene atributos considerados como verdaderos y creíbles por los usuarios en un contexto específico de uso.
Actualidad	Es el grado en el cual el dato tiene los atributos que son del período correcto en un contexto específico de uso.
Accesibilidad	Es el grado en el cual se puede acceder al dato en un contexto específico de uso.
Conformidad	Es el grado en el cual el dato tiene atributos que se adhieren a normas, convenciones o regulaciones vigentes y reglas relacionadas con la calidad de datos en un contexto específico de uso.
Confidencialidad	Es el grado en el cual el dato tiene los atributos que aseguran que éste es sólo accesible e interpretable por usuarios autorizados en un contexto específico de uso.
Portabilidad	Es el grado en el cual los datos tienen atributos que les permiten ser instalados, substituidos o movidos de un sistema a otro conservando la calidad existente, en un contexto específico de uso.

Tabla 4. Dimensiones de calidad de datos.

Los datos pueden presentar problemas en una dimensión y ser correctos en otras. Cuando esto sucede es posible que la situación que origina el problema pueda provocar inconsistencias en varias dimensiones. Es por esto que al igual que con los atributos, las dimensiones pueden agruparse en categorías para facilitar el análisis y poder encontrar patrones de problemas de calidad (Sedó, 2007).

Estas categorías y las dimensiones asociadas a ellas, están expresadas en (Bottura Filho, 2007, Calazans, 2008, Kovac et al., 1997, Wang, 1998) y se muestran a continuación en la Tabla 5.

Categoría	Dimensiones
Intrínsecas	Exactitud, objetividad, credibilidad y reputación.
Contextuales	Valor agregado, relevancia, persistencia en el tiempo, completitud y cantidad apropiada de datos.
Representación	Interpretabilidad, fácil de entender, representación consistente y representación concisa.
Accesibilidad	Accesibilidad y seguridad en el acceso.

Tabla 5. Categorías en que se agrupan las dimensiones.

En (Bottura Filho, 2007), (Wang and Strong, 1996), se conceptualizan estas categorías de la siguiente manera:

Calidad de datos intrínseca

Son los índices de calidad de la información objetiva independientemente del contexto. No solo incluye la precisión y la objetividad sino también la credibilidad y la reputación. Esto sugiere que, contrariamente a la opinión tradicional de desarrollo, los consumidores de datos también ven la credibilidad y la reputación como una parte integral de esta categoría. La exactitud y objetividad por sí sola no es suficiente para los datos que deben considerarse de alta calidad.

Calidad de datos contextuales

Son los índices de calidad de la información que pueden variar dependiendo del contexto, es decir, dependiendo de cuándo y dónde se analizan. El resultado de esta dimensión puede variar considerablemente. La agrupación de dimensiones para esta categoría demostró que la calidad de datos debe ser considerada en el contexto de la tarea en cuestión.

Calidad de datos representacional

Son los índices de calidad que hacen hincapié en las reglas del sistema y su claridad, la coherencia y la facilidad de interpretación. Incluye aspectos relacionados con el formato de los datos (representación concisa y coherente) y el significado de estos (interpretabilidad y la facilidad de comprensión). Estos dos aspectos sugieren a los consumidores de datos, concluir que los datos están bien representados, no sólo deben ser concisos y estar consistentemente representados, también deben ser interpretables y fáciles de entender.

Calidad de datos accesible

Son los índices de calidad que hacen hincapié en las reglas de los sistemas, pero solo las normas relativas a la disponibilidad y la seguridad de la información.

1.6.1 Relaciones entre dimensiones de calidad

Las dimensiones de calidad de datos no son independientes una de la otra, sino que se relacionan estrechamente entre sí. Es necesario tener en cuenta, ante un conjunto de datos, que dimensión de calidad elegir o cuál de ellas se considera de mayor valor, con respecto a otras. En (Gallo and Corrado, 2009), se mencionan las relaciones negativas y correlaciones positivas más comunes entre diferentes dimensiones de calidad de datos.

Alguna de las relaciones negativas entre dimensiones de calidad de datos se expone a continuación:

- Datos exactos, completos y consistentes podría implicar su desactualización debido al tiempo que es necesario invertir en actividades de chequeo y corrección.
- La completitud tiene mayores probabilidades de acarrear errores de inconsistencia en los datos.

Alguna de las correlaciones positivas entre dimensiones de calidad de datos se expone a continuación:

- La corrección de errores de tipo, mejoran tanto la exactitud semántica como sintáctica.
- La actualización de los datos, podría mejorar la exactitud semántica.
- La eliminación de los valores nulos, también podría mejorar la exactitud semántica.

1.7 Métricas de calidad de datos

Las métricas de calidad de datos son una forma de medir la calidad que poseen dichos datos, mediante la representación de la calidad de muchas de las dimensiones. 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 algunos casos la métrica es única y la definición teórica de una dimensión coincide con la definición operacional de la métrica correspondiente (Crippa, 2006).

1.7.1 Tipos de métricas de calidad

Las dimensiones de calidad de datos identificadas en la fase de definición de cada una de las metodologías poseen una o varias métricas asociadas y definidas en la fase de medición de dicha metodología.

La clave para la medición es el desarrollo de métricas de calidad. Estas métricas pueden ser las medidas básicas de calidad de datos tales como: la exactitud, la temporalidad, la completitud y la consistencia.

Es posible hacer una distinción entre tres tipos de métricas.

Métricas subjetivas

La evaluación se lleva a cabo por las partes relacionadas con los datos, no sólo los usuarios finales, sino también los encargados de recopilar los datos. La evaluación subjetiva se realiza generalmente a través de un cuestionario o por medio de encuestas, por tal motivo es un juicio subjetivo de la calidad de los datos, que pueden cambiar dependiendo de la persona.

Métricas objetivas

Dado un conjunto de datos en un punto específico en el tiempo, se evalúa la calidad a través de un objetivo. A menudo este tipo de métrica se implementa a través de algoritmos.

Métricas subjetivas/objetivas

Indica una métrica para medir una dimensión D que tiene la intención de medir dos dimensiones adicionales D1 y D2 que la componen y las métricas para D1 y D2 son subjetivas y objetivas. En este caso la métrica de la dimensión D se puede definir de igual forma.

En (Pipino et al., 2002), se plantea la existencia de métricas de calidad de datos basadas en tres formas funcionales como: la relación simple, la operación de min/máx y la media ponderada.

Relación simple

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, una forma de realizar la medición es el número de resultados no deseados, dividido por el total de los resultados y

restado con uno. Esta relación simple se adhiere a la convención de que uno representa el resultado más deseable y cero el resultado menos deseable.

En (Pipino et al., 2002), se sugiere utilizar la relación que muestra resultados positivos, explicando la utilidad en comparaciones longitudinales que ilustran las tendencias de mejora continua. Muchas de las dimensiones de calidad de datos, tales como: la exactitud, la completitud y la consistencia toman esta forma de medición.

Operación de min/máx

Puede ser aplicada para manejar dimensiones que requieran la agregación de múltiples indicadores de calidad de datos, mediante el cálculo de valor mínimo o máximo entre los valores normalizados de los indicadores individuales de calidad de datos.

El operador mínimo es conservador al asignar a la dimensión un valor total no mayor que el valor de su indicador de calidad de datos más débil, en este caso, evaluado y normalizado entre cero y uno. El operador máximo resulta útil en métricas mucho más complejas, aplicables a las dimensiones relacionadas con el tiempo y la accesibilidad (Pipino et al., 2002).

Media ponderada

Una alternativa al operador mínimo es el uso de la media ponderada de variables. Si una empresa, por ejemplo, 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 cero y uno y los factores de ponderación se deben añadir a uno (Pipino et al., 2002).

1.7.2 Métricas y dimensiones de calidad

La dimensión consistencia puede ser vista como: la consistencia de los mismos valores de datos a través de las tablas. Las restricciones de integridad referencial de *Codd's*, representan una instancia de este tipo de consistencia.

En (Moges et al., 2013), se expresa que la dimensión relacionada con el tiempo o dimensión temporalidad, refleja cómo los datos están actualizados con respecto a una tarea que los esté utilizando. Una métrica para este caso sería medir el tiempo de vida como el máximo entre dos

términos: cero y uno menos la relación de la actualidad entre la volatilidad, aunque dependería del contexto en el que se vaya a aplicar. La volatilidad se refiere a la longitud del tiempo en que los datos mantienen su validez (Pipino et al., 2002).

En (Fisher et al., 2009), se propone 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] 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, se utiliza una distribución de probabilidad 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).

En la literatura se muestran algunas de estas métricas y las dimensiones de calidad de datos a las que pueden estar asociadas. En (Cykana et al., 1996), (Valverde et al., 2009), se resumen algunas de ellas, mostradas a continuación en la Tabla 6.

Dimensión	Ejemplo de métrica
Precisión	Porcentaje de valores correctos en comparación con el valor real. Por ejemplo, M = masculino cuando la persona es de sexo Masculino.
Completitud	Porcentaje de campos de datos que tiene valores en ellos.
Consistencia	Porcentaje de valores coincidentes a través de tablas / campos / registros.
Temporalidad	Porcentaje de los datos disponibles en un tiempo de umbral especificado.
Unicidad	Porcentaje de registros que tienen una clave primaria única.
Validez	Porcentaje de datos que tienen valores que caen dentro de su respectivo ámbito de los valores permitidos.

Tabla 6. Métricas y dimensiones asociadas.

[1] Lempel-Ziv es un algoritmo para la comprensión de pérdidas de datos. De hecho, no es un único algoritmo, sino toda una familia de algoritmos, derivadas de los dos algoritmos propuestos por Jacob Ziv y Abraham Lempel en 1978.

Las métricas de calidad de datos pueden ser para tareas independientes o 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 las cuales incluyen reglas de organización de negocios, regulaciones de compañías y del gobierno y restricciones establecidas por administradores de base de datos, se desarrollan en contextos de aplicación específicos (Pipino et al., 2002).

Basado en (Moges et al., 2013), a continuación en la Tabla 7 se muestra el cálculo de algunas de las métricas de calidad propuestas por distintos autores, para las dimensiones de calidad de datos más comunes.

Dimensión de calidad de datos	Métricas de calidad de datos
Exactitud	$\mathbf{1} - \frac{\text{número de unidades de datos con error}}{\text{total de unidades de datos}}$ f(porcentaje de exactitud, medida de aleatoriedad con L − Z, distribución de probabilidad) acercamiento a las redes bayesianas
Completitud	$1 - \frac{\text{número de artículos incompletos}}{\text{total de artículos}}$
Consistencia	$1 - \frac{\text{proporción de violaciones de un tipo específico de consistencia}}{\text{total de verificaciones de consistencia}}$ $\frac{\text{número de valores consistentes}}{\text{total de valores}}$
Actualidad	$(\text{tiempo de entrega} - \text{tiempo de entrada}) + \text{edad}$
Volatilidad	longitud del tiempo por la cuál los datos siguen siendo válidos
Tiempo de vida	$Q_{act} = e^{-\text{disminución}(A)*\text{edad}(W*A)}$ $\max(1 - \frac{\text{actualidad}}{\text{volatilidad}}, 0)^S$

Tabla 7. Métricas para algunas dimensiones de calidad de datos.

Anotaciones importantes a tener en cuenta para un mayor entendimiento de la Tabla 7.

En el caso de la Exactitud:

- f, indica “función de”.

En el caso del Tiempo de vida:

- Qact, nivel de actualidad de los datos.
- A, atributo.
- W, 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.
- disminución, se refiere a la velocidad promedio de descenso del tiempo de vida de los valores del atributo, para el atributo bajo consideración.

Cualquiera que 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 para la calidad de datos pueden ser obtenidas a lo largo de varias dimensiones para el análisis. (Pipino et al., 2002)

1.8 Conclusiones del capítulo

Después de haber realizado una revisión bibliográfica en la literatura de los principales conceptos y definiciones que se necesitan para entender el problema planteado, se puede arribar a las siguientes conclusiones del capítulo:

Se logra un acercamiento a las posibles definiciones de calidad de datos, aclarándose que este concepto debe ser considerado de forma multifacética y relativo al uso del dato, mostrando además los problemas potenciales y las consecuencias de la mala calidad. Se identifican las principales áreas relacionadas con el término de calidad de datos y otras áreas de trabajo afines.

Se hace un esbozo de las distintas dimensiones de calidad, señalando como las más utilizadas: la exactitud, la completitud, la consistencia y las dimensiones relacionadas con el tiempo. Además se establece un conjunto de relaciones entre las propias dimensiones de calidad.

Se abordan las métricas de calidad, definiendo sus tipos, entre los cuales se encuentran las: subjetivas y las objetivas. Se precisan la asociación existente entre las métricas y las dimensiones más utilizadas, teniendo en cuenta también un muestreo del cálculo de algunas de ellas.

CAPÍTULO 2. ANÁLISIS DE LAS METODOLOGÍAS DE CALIDAD DE DATOS.

En este capítulo se realiza un análisis de las distintas metodologías de calidad de datos, mostrando los tipos de datos y dimensiones asociadas a las mismas, según las consideraciones de varios autores. Se muestra una comparación entre varias de las metodologías estudiadas y se realiza un análisis más detallado sobre la metodología TDQM.

2.1 Metodologías de evaluación de la calidad de datos

En los últimos años se han desarrollado un grupo de metodologías, que proporcionan un conjunto de pautas y técnicas que a partir de la información de entrada en relación con una realidad de interés, define un proceso racional para el uso de la información, para medir y mejorar la calidad de los datos a través de las fases dadas y los puntos de decisión.

El objetivo de las metodologías, es proporcionar una evaluación precisa y diagnosticar el estado del sistema de información con respecto a las cuestiones de calidad de datos. Los principales productos de las metodologías son la medición de la calidad de las bases de datos y los flujos de datos.

Para una mejor comprensión de las metodologías de calidad de datos existentes en la literatura, es necesario clasificarlas teniendo en cuenta varios criterios, los que están bien expresados en (Batini and Scannapieca, 2006).

Basadas en datos y orientadas a procesos

Esta clasificación se relaciona de forma general a la estrategia elegida para el proceso de mejora. Las estrategias basadas en datos tienen su punto de origen en el uso de una fuente de datos exclusiva para mejorar la calidad de los datos.

En las estrategias orientadas a procesos, se analiza el proceso de producción de los datos modificado para identificar y eliminar las causas fundamentales de los problemas de calidad.

Las metodologías de propósito general pueden adoptar ambas estrategias, con diferente profundidad según la metodología específica.

Medición y mejora

Se necesitan metodologías para la medición, la evaluación y la mejora de la calidad de los datos. Las mediciones y las actividades de mejora están estrechamente relacionadas entre sí, como consecuencia, el límite entre las metodologías de medición y mejora es en ocasiones impreciso.

De propósito general y propósito específico

Una metodología de uso general cubre un amplio espectro de fases, dimensiones y actividades, mientras que una metodología de propósito especial se centra en una actividad específica, en un dominio de datos específico o dominios específicos de aplicación.

Intra organizacional e inter organizacional

Las actividades de medición y mejora pueden referirse a un proceso o una base de datos en específico. De lo contrario, se refieren a un grupo de organizaciones que cooperan entre sí para alcanzar una meta común.

2.1.1 Resumen descriptivo de las metodologías de calidad de datos

Algunas de las metodologías existentes en la literatura se describen a continuación, mostrando una descripción general y destacando el enfoque de cada una de ellas y sus contribuciones en la evaluación de la calidad de datos y el proceso de mejora.

Metodología TDQM (Gestión Total de Calidad de Datos)

TDQM fue la primera metodología en ser publicada en la literatura, referente a la calidad de datos (Wang, 1998). Es el resultado de una investigación académica, pero se ha utilizado ampliamente como una guía para la ingeniería de datos. Su objetivo es apoyar la mejora de los procesos de calidad de datos desde el análisis de requisitos hasta la implementación. La aplicación de esta metodología requiere un considerable rediseño contextual, flexibilidad y la provisión de herramientas prácticas. La metodología constituye un problema para las grandes empresas que son dueñas de grandes bases de datos, donde una gran cantidad de personas pueden tener diferentes necesidades en el uso de los datos (Wijnhoven et al., 2007).

Metodología DWQ (Data Warehouse Quality)

Desarrollada como un proyecto de calidad dentro de los almacenes de datos (Jeusfeld et al., 1998), esta metodología estudia la relación entre los objetivos de calidad y opciones de diseño en el almacenamiento de datos. Considera la subjetividad del concepto de calidad y proporciona una clasificación de los objetivos de calidad, además, son considerados también la diversidad de los objetivos de calidad y la definición de los metadatos correspondientes.

En (Batini et al., 2009), se afirma que los metadatos de almacenamiento de datos, deben tener en cuenta tres perspectivas: una perspectiva empresarial conceptual centrada en el modelo de empresa, una perspectiva lógica centrada en el esquema del almacén de datos, y una perspectiva física que representa la capa de transporte de los datos físicos. Estas perspectivas corresponden a las tres capas tradicionales de almacenamiento de datos.

Desde la perspectiva de calidad de los datos, la metodología está caracterizada por cuatro fases: definición, evaluación, análisis y mejora. Los principales aportes de esta metodología son la clasificación de los tipos de datos y las dimensiones de calidad de software explicadas en (Batini et al., 2009).

Diseño y calidad de la administración

El primero se refiere a la capacidad del modelo para representar la información de manera adecuada y eficiente, mientras que el segundo se refiere a la forma en que el modelo evoluciona durante la operación de almacenamiento de los datos.

Calidad de ejecución de software

En esta etapa es considerada la norma ISO 9126, para las dimensiones de calidad.

Calidad de uso de datos

Se refiere a las dimensiones que caracterizan el uso y la consulta de los datos contenidos en los almacenes de datos.

Para cada dimensión, se identifican los métodos de medición que se consideren más adecuados. La lista de estos métodos, junto con el grado relevancia asociados a cada dimensión de las partes interesadas son la entrada para la etapa de medición.

En la etapa de evaluación de la calidad, el almacenamiento de la información sobre cada dimensión de calidad de datos se hace de la siguiente forma:

- Los requisitos de calidad de un intervalo de valores esperados.
- La medición de la calidad lograda.
- La métrica utilizada para calcular una medición.
- Las dependencias causales con otras dimensiones de la calidad.

La información acerca de las dependencias entre las dimensiones de calidad se utiliza para rastrear y analizar los problemas de calidad. La identificación de áreas críticas es el último paso analizado en la metodología. En esta sólo se menciona la fase de mejora, pero no contiene conocimiento constructivo acerca de cómo mejorar la calidad de un almacén de datos.

Metodología ISTAT (Oficina Nacional Italiana de Censo)

Esta metodología ha sido diseñada para uso de la administración pública italiana (Falorsi et al., 2003), con el objetivo de recopilar y mantener datos estadísticos con alta calidad de los de los ciudadanos y de las empresas del propio país. Es la metodología más aceptada a nivel académico e institucional, siendo utilizada para la detección e institucionalización de los distritos industriales en Italia.

Se caracteriza por un rico espectro de estrategias y técnicas que permiten su adaptación a varios dominios. El problema fundamental que enfrenta la metodología es la forma de garantizar la calidad de los datos integrados de múltiples bases de datos. La metodología se centra en gran medida en las normas formales ya que tiene por objeto regular las actividades de gestión de datos, de tal manera que su integración puede satisfacer los requisitos básicos de calidad.

En (Batini et al., 2009), se muestran las tres etapas fundamentales de la metodología ISTAT, junto con los flujos de información existentes entre ellos.

La etapa o fase de evaluación, se realiza inicialmente en las bases de datos por los propietarios de los datos y directivos, para detectar problemas de calidad desde una perspectiva de integración de datos. La etapa de mejora global, que se encarga de realizar la eliminación de

registros duplicados entre bases de datos nacionales y el diseño de la mejoras sobre los procesos incluyendo las decisiones a tomar, las compras o las adaptaciones a las soluciones existentes. Las actividades de mejora que requieran la cooperación de múltiples administraciones están típicamente orientadas a procesos, ya que dirigen los flujos de datos que se intercambian durante la ejecución de las actividades específicas. Estas actividades se planifican de forma centralizada y coordinada.

La metodología ISTAT proporciona una variedad de técnicas estadísticas sencillas pero eficaces para la medición de la calidad. También proporciona herramientas para las actividades de limpieza de los datos. La metodología apoya la normalización de los formatos de datos y su expresión en un esquema XML (Lenguaje de Marcas Extensible) común, que permite la integración de las bases de datos de los administradores locales. Los datos intercambiados entre diferentes administraciones son rediseñados usando una arquitectura de software orientada a eventos, basado en mecanismos de publicación y suscripción.

Metodología AIMQ (Metodología para la Evaluación de la Calidad de la Información)

Es una metodología para la evaluación de la calidad de la información centrada en la evaluación comparativa (Lee et al., 2002), además es una técnica objetiva e independiente de dominio para la evaluación de la calidad.

El fundamento de AIMQ se basa en un modelo llamado PSP/IQ y un conjunto de dimensiones que cubren aspectos de calidad de información las cuales son importantes para los consumidores de la información, (Lidiansa, 2014). Las publicaciones que describen AIMQ se centran principalmente en las actividades de evaluación, mientras que las directrices, técnicas y herramientas para actividades de mejora no se proporcionan, ni se menciona ninguna descripción de la base de datos de evaluación comparativa que se requiere para la aplicación de la metodología (Batini et al., 2009).

La primera componente de esta metodología es una matriz cuadrada (en este caso 2x2) llamada también modelo o marco de lo que significa calidad de la información para los consumidores y gestores de información (Lee et al., 2002). El modelo cuenta con cuatro cuadrantes, dependiendo si la información es considerada un producto o un servicio y de si las

mejoras pueden ser evaluadas con una especificación normal o por las expectativas de los clientes, conocido también como: PSP/IQ.

El segundo componente es denominado evaluación de la calidad de la información, esta se realiza mediante la aplicación de un cuestionario para medir la calidad de la información utilizando las dimensiones de calidad. Siendo importantes para los consumidores y gestores de información, esta componente puede ser aplicada también para evaluar la calidad de la información en las organizaciones (Lee et al., 2002).

El tercer componente de la metodología, expresada en (Calazans, 2008), consta de dos técnicas de análisis para la interpretación de las evaluaciones tomadas en los cuestionarios. La primera técnica compara la calidad que posee la información en la organización y utilizarla como un punto de referencia. La segunda técnica mide las distancias entre las evaluaciones de los diferentes actores de un sistema de producción de la información.

Estas dos técnicas ayudan a las organizaciones en centrar sus esfuerzos de mejora en el análisis de sus evaluaciones de información (Lee et al., 2002).

Modelo PSP/IQ

El fundamento de la metodología AIMQ es un modelo y un conjunto de dimensiones de calidad que cubren aspectos de calidad de la información que son importantes para los consumidores de dicha información. El modelo organiza las dimensiones con el objetivo de tomar decisiones significativas en vista de mejorar la calidad de los datos, desarrollando estas dimensiones desde la perspectiva de los consumidores de la información.

El modelo sitúa las dimensiones en cuatro cuadrantes: la información, confiable, útil e información utilizable (Lee et al., 2002). Estos cuatro cuadrantes representan aspectos de calidad de la información que son relevantes para las decisiones de mejora de la calidad.

	Conformidad de las especificaciones	Expectativas del cliente
Calidad del producto	La información. Dimensiones de calidad (representación, completitud, consistencia).	Información útil (la información proporcionada cumple con el trabajo las necesidades del usuario). Dimensiones de calidad (relevancia, interpretabilidad, objetividad).

Calidad de los servicios	Información confiable. Dimensiones de calidad (dimensiones de tiempo).	Información útil. Dimensiones de calidad (credibilidad, accesibilidad).

Tabla 8. Cuadrantes del modelo PSP/IQ.

La metodología AIMQ en su conjunto proporciona una herramienta práctica para las organizaciones en el área de la calidad de la información. Se ha aplicado en diversos contextos organizacionales, tales como las industrias financieras, de salud y la fabricación (Lee et al., 2002).

Metodología CIHI (Instituto Canadiense de Información Sanitaria)

Esta metodología ha puesto en marcha un método para evaluar y mejorar la calidad de datos de información de la salud en el mencionado instituto sanitario (Long and Seko, 2005). En el escenario CIHI, el principal problema es el tamaño de las bases de datos y su heterogeneidad. La metodología apoya la selección de un subconjunto de datos para enfocar la fase de evaluación de la calidad. También propone un amplio conjunto de criterios de calidad para evaluar la heterogeneidad.

En (Batini et al., 2009), se expresa que la estrategia de esta metodología propone un enfoque en dos fases. En la primera se define un marco de calidad de datos, y en la segunda un análisis en profundidad a los datos de acceso frecuente. El marco de calidad de datos se define en tres pasos:

- La normalización de la información de calidad de datos.
- El desarrollo de una estrategia común para la evaluación de calidad de los datos.
- La definición de un proceso de trabajo para la gestión de datos CIHI que identifica las prioridades de calidad de datos e implementa procedimientos de mejora continua de los datos.

El análisis de los datos de uso más frecuente se realiza en tres etapas: análisis de la calidad de datos, evaluación y documentación. En esa última se informan los problemas de calidad detectados por el análisis y la evaluación de la calidad de los datos.

Metodología DQA (Evaluación de Calidad de Datos)

Esta metodología de calidad de datos ha sido diseñada para proporcionar los principios generales que guían la definición de métricas de calidad de datos (Pipino et al., 2002). En la literatura, las métricas de calidad de datos se definen en su mayoría para resolver problemas específicos, por tanto, dependen del escenario en que se use.

En (Batini et al., 2009), se sintetiza que la metodología distingue entre indicadores subjetivos y objetivos de calidad. Las métricas subjetivas miden las percepciones, necesidades y experiencias de los grupos de interés. Las métricas objetivas se clasifican en tareas independientes y tareas dependientes. El primero evalúa la calidad de los datos sin conocimiento contextual de la aplicación, mientras que el segundo se define para contextos de aplicación específicos e incluyen reglas de negocio, normativas de la empresa y del gobierno además de las restricciones previstas por la administración de la base de datos. Ambos indicadores se dividen en tres clases: la relación simple, la operación de min/máx y la media ponderada, antes expuesta.

Metodología IQM (Medición de Información de calidad)

Proporciona un marco de calidad de la información a la medida de datos de la Web. En particular IQM ayuda a la selección basada en la calidad y la personalización de las herramientas que utilizan los *webmasters* en la creación, gestión y mantenimiento de sitios web (Eppler and Muenzenmayer, 2002).

Esta metodología para la calidad de los datos de las organizaciones, se basa en niveles que miden el estado actual, plantean planes de acción y establecen el grado de madurez alcanzado por las empresas en su objetivo de garantizar información de calidad, desde sus datos con calidad. La creación de este modelo parte de la adaptación de la metodología *Qinfo*, la cual fue elaborada por un grupo de investigación en el año 2005 y que estableció los fundamentos sobre los que se soporta el modelo actual (Rosales and Herrera, 2012).

En (Rosales and Herrera, 2012), se plante que la metodología IQM se fundamenta en la creación de un modelo de madurez que a partir de la identificación y valoración de las prácticas empresariales que están directamente relacionados con el tratamiento de los datos, se establecen y recomiendan técnicas, herramientas o procedimientos que ayuden a fortalecer y

mejorar las condiciones de los datos que los lleve a un sistema efectivo de aseguramiento de calidad y que permita mantenerla en el tiempo.

Además tiene como otros objetivos apoyar a las organizaciones en el fortalecimiento de sus prácticas y tecnologías para lograr sistemas de información consistentes, soportados en datos de alta calidad. Integrar la calidad de los datos a la gestión de la organización, clave para el cumplimiento de metas, generación de oportunidades, mejorar indicadores financieros y de eficiencias. Iniciar un proceso de fortalecimiento de la calidad de sus sistemas información a través de recomendaciones y propuestas de acciones de inmediato, corto y mediano plazo (Rosales and Herrera, 2012). Para alcanzar esto, la metodología identifica la condición actual de la empresa; identifica los riesgos asociados a la calidad de los datos; identifica las brechas tecnológicas que debiliten las condiciones sobre los datos que soportan los procesos organizacionales y establece un plan de acción para evolucionar y gestiona sus resultados.

La metodología proporciona dos conjuntos de directrices: el marco de la calidad de información que define los criterios de calidad, y el plan de acción que explica cómo realizar las mediciones de calidad.

El modelo de madurez IQM es un modelo adaptado de las mejores prácticas de medición de las capacidades empresariales, enfocado al análisis y valoración del tratamiento y administración de los datos (Gama et al., 2013). Examina un conjunto de prácticas y procesos que se han agrupado en cuatro dimensiones o perspectivas de la organización y sobre las cuales se considera que deben ser incorporadas técnicas, herramientas y mecanismos para que la estructura de datos, que se maneja hacia el interior de la organización, cuente con los elementos de calidad requeridos.

En (Gama et al., 2013), se expresa que la metodología en comparación con otros modelos de madurez cuenta con los siguientes detalles.

- Aplica un enfoque sistemático.
- Emplea un proceso progresivo de calidad de los datos.
- Formula recomendaciones para su aplicación práctica y efectiva.
- Emplea un enfoque estratégico del modelo y lo vincula con la gestión del conocimiento.

- Vincula diferentes dimensiones o perspectivas de las organizaciones.
- Propone mejoras de procesos en operación y alternativas de implementación de productos, además de utilizar métodos que logran altos impactos en la efectividad operativa, técnica y financiera.

La metodología IQM de calidad de datos se fundamenta sobre cuatro componentes:

- Los referentes de calidad de datos.
- El ciclo de vida del dato.
- Las perspectivas organizacionales.
- Los niveles de madurez.

Los referentes de calidad de datos

Se establecen para determinar el tipo de inconsistencia que pueden presentar los datos a partir de los factores de vulnerabilidad identificados. Estos referentes pueden ser de tipo: dato exacto; dato integral; dato consistente; dato completo.

Ciclo de vida del dato

Existen cuatro estados o fases que determinan el ciclo de vida del dato en la metodología, que son: el origen, la administración, la transformación y el uso, al contemplar la información como un producto de procesamiento de datos.

Perspectivas organizacionales

En la metodología se afirma que la calidad de los datos se alcanza a través de la ejecución de buenas prácticas en todo el ciclo de vida del dato (Gama et al., 2013). Para mantener esto IQM se enfoca en los sistemas de la información estratégica, entendida como el grupo de componentes que tiene mayor impacto en la calidad de la información y formada por dos enfoques: los datos evaluados con los referentes de calidad definidos y las perspectivas organizacionales, esta última formada por la dirección de los procesos de su ciclo de vida, la definición y ejecución de estos, la cultura organizacional orientada a la calidad, la arquitectura y la seguridad. A continuación se muestra en la Tabla 9 (Gama et al., 2013), las perspectivas organizacionales del modelo IQM.

Perspectivas	
Dirección	Prácticas que determinan el nivel de preparación sostenible de una organización para generar información de calidad para la toma de decisiones eficientes, evidenciado a través de directrices enunciadas, oficiales, implementadas y auditadas que estimulan la calidad de los datos y su sistema de indicadores.
Cultura organizacional	Es el grado de apropiación e incorporación de la calidad de los datos en la organización, vista a través de la difusión de las políticas asociadas y su aplicación. En el proceso, esta perspectiva está estrechamente ligada a los factores de la perspectiva de Dirección, realizando una incorporación de las directrices y de la gestión de la calidad de los datos.
Trazabilidad	Factores empresariales que permiten determinar el nivel de la presencia de buenas prácticas en los procedimientos de validación, medición, seguridad y mejoramiento continuo, tal que se garantice la calidad de los datos durante todo el ciclo de vida.
Arquitectura	Es el conjunto de características de los insumos de tecnología utilizados para generar la información estratégica de la organización, compuesto por la arquitectura de datos; de sistemas de información y la ingeniería aplicada, cuyo nivel de calidad influye en la capacidad para asegurar la calidad del dato.

Tabla 9. Perspectivas organizacionales del modelo de madurez IQM.

Niveles de madurez del modelo IQM

En (Rosales and Herrera, 2012), se especifican tres niveles de madurez para alcanzar altos estándares.

Básico: Los sistemas de información están identificados, implementados y verificados.

Medio: Los elementos que afectan la calidad de los datos se encuentran controlados y gestionados.

Alto: Las actividades están estables y se aplica mejoramiento continuo.

Cada nivel tiene una serie de requerimientos para los elementos más influyentes sobre la calidad de los datos los cuales son: la información estratégica y operativa; los datos trazados verificables y seguros; la medición de la calidad de los datos; los procesos de soporte tecnológico; los datos con tratamiento técnico y la calidad conocida y apropiada.

La metodología IQM es una herramienta capaz de identificar las condiciones actuales de la empresa; los riesgos asociados a la calidad de los datos y las brechas tecnológicas que debilitan las condiciones sobre los datos que soportan los procesos organizacionales.

Además presta un beneficio importante al reducir los costos operativos que produce la mala calidad; ofrece productos oportunos y mejor adaptados a los clientes; evita la perdida de información crítica o estratégica y mejora la calidad; atención; servicio y productividad al incorporar mejores prácticas en la cultura organizacional.

Metodología AMEQ (Basado en actividades de medición y evaluación de la calidad del producto información)

El objetivo de esta metodología es proporcionar una base rigurosa para la calidad del producto de información; evaluación y mejora en el cumplimiento de las metas de la organización (Su and Jin, 2006).

La metodología es específica para la evaluación de calidad de los datos en las empresas de fabricación, donde la información del producto representa el principal componente de las bases de datos operativas. En las empresas manufactureras, la asociación entre los procesos de información del producto y de producción es sencilla y relativamente estándar en las empresas. El esquema de las bases de datos de los productos también es similar en diferentes organizaciones. La metodología proporciona un enfoque y directrices metodológicas para modelar tanto la información como los procesos de producción relacionados.

La metodología AMEQ consta de cinco fases para los procesos de medición y mejora. La primera fase mide la adecuación cultural de una organización, utilizando la red de Madurez de

Información de Gestión de Calidad y utiliza una plantilla para realizar entrevistas para puestos clave de dirección. La segunda fase especifica el producto de la información (Batini et al., 2009). Además cada producto de información se relaciona con un proceso de negocio correspondiente, siguiendo el modelo por medio de un enfoque orientado a objetos.

En esta metodología de calidad de datos, se modelan ocho tipos de objetos: los recursos humanos; los recursos de información; las actividades empresariales; las aportaciones de recursos; los procesos de recursos; las salidas de recursos; las medidas de desempeño y las metas de la empresa. En esta fase se produce también un modelo de métodos de medición.

La tercera fase se centra en la actividad de medición. En la cuarta fase se investigan las causas fundamentales de los problemas mediante el análisis de las dimensiones de la calidad que han recibido una puntuación baja. Por último, se inicia la etapa o fase de mejora.

Metodología COLDQ (Costo-Efecto de la Baja Calidad de Datos)

El objetivo fundamental de la metodología COLDQ es proporcionar un sistema de puntuación de calidad de datos de apoyo a la evaluación de la relación costo-efecto de la baja calidad de los datos (Loshin, 2001). La metodología proporciona una clasificación detallada de los costos y beneficios. Los beneficios directos se pueden obtener evitando los costos de mala calidad debido a la adopción de técnicas de mejora.

Teniendo como propósito obtener una evaluación cuantitativa de la medida en que los procesos de negocio se ven afectados por la mala información, la metodología propone fases interrelacionadas para evaluar el costo-efecto de la baja calidad de los datos.

En la primera fase de la metodología, el contexto empresarial se modela mediante la identificación de dos modelos de flujo de datos: el flujo de datos estratégicos que se utiliza para la toma de decisiones y el flujo de datos operativos que se utiliza para el procesamiento de datos (Batini et al., 2009). Ambos modelos representan un conjunto de etapas de procesamiento que describen el flujo de información desde el suministro de datos para el consumo de dichos datos.

Sobre la base de estos modelos, se llevan a cabo los análisis objetivos y subjetivos del contexto empresarial. Los errores identificados se atribuyen a las actividades defectuosas en

los modelos estratégicos y operativos del propio contexto. Esta asociación entre los errores y actividades proporciona la base para las evaluaciones de costos.

En (Batini et al., 2009), se expresa que la metodología proporciona una clasificación exhaustiva y valiosa de los impactos económicos operativos, tácticos y estratégicos que tienen que ser considerados. Cada clase de costos se le asigna un valor económico basado en el conocimiento contextual. Los costos representan la entrada a la fase de mejora final. La metodología de calidad de datos COLDQ soporta análisis costo-beneficio, mediante la evaluación y la suma del costo de los proyectos de mejora de calidad.

Metodología DaQuinsCIS (Calidad de datos en los sistemas de información de una cooperativa)

Esta metodología trabaja en dirección de resolver los problemas de calidad de datos en los sistemas de información cooperativos (Scannapieco et al., 2004). La arquitectura DaQuinCIS ha sido diseñada para gestionar la calidad de datos en contextos de cooperación, a fin de evitar la propagación de datos de baja calidad y para explotar la replicación de datos para la mejora de la calidad global de los datos.

El componente principal es el *broker* de calidad de datos. Este posee dos funcionalidades principales: la intermediación de calidad, que permite a los usuarios seleccionar los datos en función de su calidad y el mejoramiento de la calidad, que se difunde a los mejores ejemplares de calidad de los datos en el sistema (Milano et al., 2006). Atendiendo a la funcionalidad de la intermediación de calidad, el *broker* de calidad de datos es, en esencia, un sistema de integración de datos que permite el acceso a los datos de mejor calidad disponibles sin tener que saber dónde se almacenan estos datos.

2.1.2 Metodologías y tipos de datos

Como explicamos en el epígrafe: 1.2, los tipos de datos influyen en las dimensiones de calidad y en las técnicas de evaluación, además se asocian a una metodología.

En (Batini et al., 2009), se expone que un grupo de las metodologías abordan los datos estructurados, mientras que solo unas pocas tratan los datos semiestructurados.

Algunas de las metodologías y los tipos de datos asociados se resumen en la Tabla 10.

Metodología	Datos estructurados	Datos semiestructurados
TDQM	X	x
DWQ	X	
TIQM	X	Considerado implícitamente
AIMQ	X	Considerado implícitamente
CIHI	X	x
DQA	X	
IQM	X	x
ISTAT	X	x
AMEQ	X	Considerado implícitamente
COLDQ	X	Considerado implícitamente
DaQuinCIS	X	x
QAFD	X	
CDQ	X	x

Tabla 10. Metodologías y tipos de datos.

2.1.3 Metodologías y dimensiones asociadas

En (Batini et al., 2009), se muestran las dimensiones de calidad a juicio de las metodologías, una dimensión se asocia con una metodología si esta proporciona una definición correspondiente. La gran variedad de dimensiones definidas en las metodologías confirman la complejidad del concepto de calidad de datos.

Algunas de las metodologías y las dimensiones que se le asocian se resumen en la Tabla 11.

Metodología	Dimensión de calidad
TDQM	Accesibilidad, exactitud, credibilidad, consistencia, interpretabilidad, comprensibilidad, completitud y dimensiones relacionadas con el tiempo.
DWQ	Exactitud, integridad, trazabilidad, interpretabilidad, accesibilidad y coherencia.
TIQM	Consistencia, integridad, precisión, exactitud, precisión, accesibilidad, oportunidad y claridad contextual.

AIMQ	Accesibilidad, adecuación, credibilidad, integridad, consistencia, representación, interpretabilidad, objetividad, relevancia, comprensibilidad.
CIHI	Exactitud, actualidad, comparabilidad y relevancia.
DQA	Accesibilidad, credibilidad, integridad, consistencia, representación, relevancia e interpretabilidad.
IQM	Accesibilidad, coherencia, oportunidad, aplicabilidad y precisión, trazabilidad.
ISTAT	Exactitud, integridad y consistencia.
AMEQ	Interpretabilidad, exactitud, credibilidad, reputación y accesibilidad.
COLDQ	Precisión, coherencia, puntualidad y consistencia.
DaQuinCIS	Exactitud, integridad, consistencia y confiabilidad.
QAFD	Exactitud, coherencia e integridad.
CDQ	Precisión, exactitud, integridad, consistencia, puntualidad, volatilidad y accesibilidad.

Tabla 11. Metodologías y dimensiones de calidad.

2.2 Comparación de metodologías de calidad

En (Batini et al., 2009), se tienen en cuenta varias perspectivas que se pueden utilizar para analizar y comparar metodologías de calidad de datos. Por esta razón se dice que las metodologías difieren una de otra por la forma en que se consideran estos puntos de vista.

Algunas de las perspectivas de análisis que se tiene en cuenta en las metodologías de calidad, se manifiestan en la Tabla 12, en la cual se muestra una comparación de algunas de las metodologías antes estudiadas:

Metodología	Aspectos de comparación		
	Tipos de datos	Sistemas de información	Fases o pasos que componen la metodología
TDQM	Estructurados, semiestructurados	Monolítico, distribuido	Fase 1: definición Fase 2: medición Fase 3: análisis Fase 4: mejora (ciclo continuo)
DWQ	Estructurados	Almacenes de datos	Fase 1. definición Fase 2. medición

			Fase 3. análisis Fase 4. mejora
AIMQ	Estructurados	Monolítico, distribuido	Fase 1. medición Fase 2. análisis e interpretación de mejora
CIHI	Estructurados, semiestructurados	Monolítico, distribuido	Fase 1. análisis, Fase 2. medición Fase 3. documentación
DQA	Estructurados	Monolítico, distribuido	Fase 1. medición subjetiva y objetiva Fase 2. comparación Fase 3. mejora
IQM	Estructurados, semiestructurados	Web	Fase 1. planificación de la valoración Fase 2. configuración de la valoración Fase 3. medición Fase 4. continuación de actividades
ISTAT	Estructurados, semiestructurados	Monolítico, distribuido, cooperativo	Fase 1. valoración Fase 2. mejora global Fase 3. mejora de la calidad de datos interna o mejora inter administrativa
AMEQ	Estructurados	Monolítico	Fase 1. establecer el ambiente del sistema de información Fase 2. definición Fase 3. medición Fase 4. análisis Fase 5.mejora
COLDQ	Estructurados	Monolítico	Fase 1. información Fase 2. análisis Fase 3. separar datos erróneos Fase 4. identificar dominios de impactos Fase 5. evaluación de costos,

Tabla 12. Comparación entre metodologías de calidad de datos.

La comparación realizada anteriormente, indica claramente que las metodologías tienden a centrarse en un subconjunto de cuestiones de calidad de datos. Las grandes diferencias de enfoque a través de las mismas pueden ser reconocidas en cuatro categorías (Gallo and Corrado, 2009).

- Metodologías completas: proporcionan apoyo a las fases de evaluación y de mejora y ayudan a afrontar las cuestiones técnicas y económicas.
- Metodología de auditoría: se centran en la fase de evaluación y proporcionar un apoyo limitado a la fase de mejora.
- Metodologías operacionales: se centran en los aspectos técnicos de las fases tanto de la evaluación y de mejora, pero no se ocupan de las cuestiones económicas.
- Metodologías económicas: se centran en la evaluación de los costos.

Después de lo antes expuesto en este capítulo, se decide trabajar y profundizar más en la metodología TDQM, al ser esta una de las primeras metodologías en surgir en el tema de la calidad de datos. Además de ofrecer un apoyo completo a las fases de evaluación y mejora de la calidad, abarcando los datos de tipo estructurados y semiestructurados.

2.3 Análisis descriptivo de la metodología de calidad de datos TDQM

A continuación se realiza un análisis de la metodología TDQM, mostrando una descripción detallada de la misma, destacándose su enfoque de trabajo y su contribución en la evaluación de la calidad de datos y proceso de mejora.

Metodología TDQM

TDQM fue la primera metodología en ser publicada en la literatura, referente a la calidad de datos (Wang, 1998). El objetivo de esta metodología es generar productos de información de buena calidad para los consumidores de información.

En (Daza et al., 2012), se aborda la metodología TDQM, la cual considera cuatro fases para la gestión de los productos de información, las cuales implementan una mejora continua de la calidad, estas fases son: ***definición***, ***medición***, ***análisis*** y ***mejora***. Se ejecutan iterativamente, constituyendo así un ciclo el cual es 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 dentro del tema de calidad de datos y en especial de la metodología.

Fase de definición

La fase de definición incluye la identificación de las dimensiones de calidad de datos y los requisitos relacionados. Además es necesario conceptualizar los siguientes aspectos expuestos en (Guzmán and de Llergo, 2013).

1. Las características del producto de información. Se conceptualiza el producto de información de acuerdo a la funcionalidad para el consumidor de la información, es decir, la información de los clientes que se necesita para realizar las tareas. También se identifican los consumidores y funcionalidades del sistema capturando toda la información que estas personas crean necesarias para realizar las mismas.

   Debido a que no todos los productos de información y atributos pueden ser evaluados, es útil centrarse en el producto de información o atributo más importante, que pueden ser aquellos cuyos problemas de calidad tienen mayor impacto (Wijnhoven et al., 2007).
2. Los requerimientos de calidad en la información, una vez definidas las características del producto, se definen intentando abarcar los intereses de los usuarios, proveedores, procesadores y administradores de la información.
3. El sistema de producción de la información, es decir, el conjunto de personas, equipos y programas computacionales involucrados en la transformación y diseminación de la información de la organización.

Fase de medición

En la fase de medición se especifican las métricas asociadas a las dimensiones de calidad identificadas en la fase anterior, que proporcionan información para la gestión de calidad de los datos.

En este punto se debe establecer qué tipo de mediciones se realizarán sobre el conjunto de datos. Las mediciones deben determinar los aspectos de calidad en relación a la satisfacción del cliente. Las mediciones deben realizarse en el lugar correcto, es decir, en el segmento de los datos que contenga la información que regularmente se utiliza para los procesos del negocio. Además se deben identificar los parámetros de medición a utilizar, los cuales serían

los componentes más importantes de las dimensiones de calidad de datos requeridas para los sistemas de información que se estén analizando (Sedó, 2007).

En (Pipino et al., 2002), se identifican tres formas de medición de las dimensiones de calidad de datos, también expuestas anteriormente:

- Relación simple.
- Operación de mínimo (min) y máximo (máx.).
- Media ponderada.

Una vez que se determinan las métricas, se puede llevar a cabo el proceso de medición. Si hay, por ejemplo, 20.000 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 and McCabe, 1989).

En (Guzmán and de Llergo, 2013), se expresa, la existencia en la literatura de diversas opiniones sobre las dimensiones de la información a las que deben aplicarse las métricas para evaluar su calidad, algunas de las más comunes son la exactitud; la cantidad de información; la facilidad de manipulación y entendimiento.

Fase de análisis

La fase de análisis identifica la raíz de los problemas de calidad de datos, los métodos y técnicas para esto varían en complejidad. Es en esta etapa donde las organizaciones deben investigar las causas de los problemas de calidad en la información.

En (Guzmán and de Llergo, 2013), se resume que los métodos para llevar a cabo esta fase son muy variados y pueden ser tanto simples, es decir, entrevistarse con los proveedores de la información, como complejos, en este último caso con el uso de herramientas de control estadístico de procesos, reconocimiento de patrones o identificación de observaciones atípicas.

Además son necesarios determinar los procesos a evaluar. El objetivo de este punto es identificar para cada proceso, en el cual los datos deben ser evaluados, con la finalidad de llevar a cabo una medición eficaz (Sedó, 2007). Es necesario evaluar la calidad de la información midiendo los datos contra las dimensiones de calidad, con el fin de determinar su nivel de fiabilidad y así poder establecer el tipo y grado de inconsistencias que se presenten.

En este punto son útiles las funciones de *profiling* de las herramientas de calidad de los datos, las cuales permiten conocer si los datos son conformes de acuerdo con las reglas del negocio, si los valores registrados son válidos y si se encuentran dentro de los límites razonables dentro del contexto del negocio. Dentro del proceso de interpretación y reporte sobre la información de calidad, es necesario comunicar el estado de calidad de los datos, de manera que se identifiquen los procesos que requieren mejoras basados en el impacto de los defectos que se han detectado.

Fase de mejora

Por último, en la fase de mejora del producto de información, las organizaciones deben identificar los elementos en los que es necesario incidir para motivar el cambio y eliminar los problemas de la calidad en la información, diseñando para esto actividades de mejora de la calidad, identificando las áreas clave para las mejoras.

De acuerdo con estos principios TDQM propone un lenguaje para la descripción de los procesos de producción de información, llamados IP-MAP (Información Mapa Producto). Siendo extendido hacia UML (Lenguaje Unificado de Modelado) para apoyar al diseño organizacional, IP-MAP es el único idioma utilizado para obtener información de modelado de procesos representando un estándar (Shankaranarayanan and Wang, 2007).

Para aplicar esta metodología es fundamental que la información sea tratada como un producto que se mueve a través de un sistema de fabricación de información, muy similar a un producto físico, pero con las ventajas y limitaciones propias de la naturaleza de un producto de información. TDQM ha demostrado ser eficaz para mejorar los productos de información, sobre todo cuando hay un fuerte compromiso desde el nivel gerencial expresado en su política organizacional (Caro et al., 2013).

Existen en la literatura tres herramientas para mejorar la calidad de datos, las cuales se han adaptado a la metodología (Daza et al., 2012).

Data profiling (Perfilado de datos)

El perfilado de datos es el proceso de obtener una comprensión de los datos existentes en relación con las especificaciones de calidad (Geiger, 2004). Además es el proceso de examinar

los datos que existen en las fuentes de origen de una organización y recopilar estadísticas e información sobre los mismos. El propósito de dichas estadísticas es:

- Determinar qué datos pueden ser usados para otros propósitos.
- Conseguir métricas de calidad de datos que incluyen si los datos cumplen los estándares de la organización.
- Reducir el riesgo de integrar información a nuevas aplicaciones dado que se conoce su estado.
- Realizar seguimiento de la calidad de datos.
- Entender problemas derivados de los datos en proyectos que hagan uso intensivo de los mismos.

Data cleansing (Limpieza de datos)

La limpieza de datos es el proceso de detectar y corregir datos corruptos, incoherentes o erróneos. Después del proceso, la información será consistente con otros conjuntos similares de datos. Este proceso permite detectar entradas duplicadas, incompletas u otros problemas y establecer reglas para corregirlas. El objetivo no es borrar información sino mejorar la calidad de los datos construyendo un proceso de mejora continua.

Data auditing (Auditoria de datos)

Se define como la aplicación de datos de minería y algoritmos para medir la calidad de los datos. Es el proceso de gestionar cómo los datos se ajustan a los propósitos definidos por la organización. Se establecen políticas para gestionar los criterios de datos para la empresa.

En (Luebbers et al., 2003), se expresa que la auditoría de datos se puede dividir en dos sub tareas, la primera de ellas es la inducción de una descripción estructural y la segunda la comprobación de los datos que marca desviaciones como posibles errores en los datos y genera correcciones probables. Ambas tareas se pueden ejecutar de forma asincrónica,

El poder de la metodología TDQM se deriva de la investigación multidisciplinaria acumulada y la práctica en una amplia gama de organizaciones, quienes deben aprovechar todo el potencial de sus datos a fin de obtener una ventaja competitiva y alcanzar los objetivos estratégicos.

2.4 Conclusiones del capítulo

Se hace un análisis de las distintas metodologías de evaluación de la calidad de datos existentes en la literatura, abordándose algunas como: la metodología DWQ, desarrollada como un proyecto de calidad dentro de los almacenes de datos, la italiana ISTAT diseñada para uso de la administración pública del propio país y la metodología IQM. Se abordan los tipos de datos asociados y las dimensiones que por lo general son las más utilizadas dentro de cada una de las metodologías de calidad.

Se plantean algunos aspectos importantes a tener en cuenta a la hora de comparar metodologías de calidad y se determinó escoger la metodología TDQM para aplicarla a un caso de estudio.

CAPÍTULO 3. APLICACIÓN DE LA METODOLOGÍA TDQM AL SIGENU

En este capítulo se aplica la metodología TDQM a la base de datos el sistema SIGENU (Sistema de Gestión de la Nueva Universidad) de la UCLV (Universidad Central "Marta Abreu" de Las Villas).

Se describen los resultados de las fases de definición y medición que plantea la metodología TDQM, proponiéndose las dimensiones de calidad de datos que serán analizadas y las métricas que se asocian a las mismas para efectuar las mediciones.

3.1 Descripción del SIGENU

El SIGENU es un sistema concebido para la gestión de la información sobre los estudiantes en las universidades cubanas. Está diseñado como una aplicación web, cuya finalidad es permitir la gestión de toda la información académica vinculada con la educación superior en Cuba. En correspondencia con su carácter nacional y la gran diversidad de sistemas de enseñanza superior con que cuenta la universidad cubana, este sistema ha sido concebido de manera tal que sea capaz de brindar gran seguridad e integridad de la información, y a la vez, ser tan flexible que permita ser adaptado a todos los centros de educación superior del país con sus diversas particularidades y distintas maneras de realizar determinados procedimientos.

En la UCLV el sistema se encuentra adaptado para trabajar en las diferentes modalidades de estudio: el curso regular diurno, el curso por encuentro y el curso de educación a distancia. Manteniendo actualizada la información de las facultades, las carreras, los planes de estudio y los estudiantes. Existe un módulo que controla el postgrado, pero que no será objeto de análisis en el presente trabajo.

La información en este sistema está organizada en una base de datos operacional y un mercado de datos sobre el gestor de datos PostgreSQL. El mercado de datos tiene el propósito específico de gestionar la información de matrículas, bajas y egresados y sus datos se cargan y transforman desde la base de datos operacional.

El PostgreSQL es un gestor de datos objeto-relacional, bajo licencia BSD (distribución de software *Berkeley*) que tiene menos restricciones en comparación con otras como la GPL (Licencia Pública General de GNU) estando muy cercana al dominio público. La licencia BSD

al contrario que la GPL permite el uso del código fuente en software no libre. Este sistema de gestión de bases de datos de código abierto más avanzado del mundo y en sus últimas versiones posee muchas características que sólo se podían ver en productos comerciales de alto calibre (Reyes, 2012).

El servidor de la aplicación del SIGENU es un servidor JBoss de aplicaciones Java EE de código abierto implementado en Java puro. El JBoss puede ser utilizado en cualquier sistema operativo para el que esté disponible la máquina virtual de Java. Es el primer servidor de aplicaciones de código abierto, preparado para la producción y disponible en el mercado, ofreciendo una plataforma de alto rendimiento para aplicaciones de *e-business*. Combinando la arquitectura orientada a servicios SOA (Arquitectura Orientada a Servicios), con la licencia GNU de código abierto.

3.2 Aplicación de la metodología TDQM

Como se explica en el capítulo 2, la aplicación de la metodología TDQM está regida por las fases que en ella se definen.

3.2.1 La fase de definición

En la fase de definición se establecen:

- Las características del producto de información,
- Los consumidores y
- Las funcionalidades del sistema

A continuación se describen cada uno de estos elementos.

Características del producto de información

En este trabajo la aplicación de la metodología se realiza sobre la base de datos operacional del SIGENU, que es la base principal del sistema.

La información, en la base de datos, está estructurada en 126 tablas que recogen toda la información referente a los estudiantes, las facultades, carreras y tipos de cursos.

De estas tablas, las de mayor importancia en el sistema, debido a su uso, son las que almacenan la información de los estudiantes, la que se encuentra fundamentalmente en la tabla ***student*** y en algunas otras que se describen más adelante.

Tabla: *student*

Esta tabla mantiene la información relacionada con los estudiantes, guardando en 35 atributos los datos generales tales como: nombre y apellidos, dirección particular, fecha de nacimiento, email, teléfono, fecha de ingreso a la educación superior, índice académico en el momento del ingreso, fecha en que es registrado la base de datos, foto, carrera, procedencia, sexo, etc. De estos 35 campos 17 son llaves ajenas a otras tablas.

Los datos utilizados en la tabla *student*, son en su mayoría de tipo *character varying (1024)*, excepto los pertenecientes a la fecha de nacimiento, fecha de ingreso a la educación superior, registro en el sistema, que son de tipo *date y timestamp*, el índice académico que es de tipo *real*, los atributos *reoffer* y *option* de tipo boolean y el atributo block que es de tipo integer.

En la tabla *student* están definidas las siguientes restricciones:

- Restricción de clave primaria en el atributo id_student
- Restricciones de llaves ajenas sobre los atributos: academic_situation_fk, career_fk, country_fk, course_type_fk, current_student_status_fk, entry_source_fk, faculty_fk, marital_status_fk, orphan_fk, scholastic_origin_fk, sex_fk, skin_color_fk, student_type_fk, student_regimen_fk, town_fk, town_university_fk, politic_org_fk.
- Restricción de no permitir nulos en los atributos que son llaves ajenas excepto career_fk, town_fk, student_status_fk, faculty_fk, town_university_fk y current_student_status_fk
- Restricción de valor por defecto en la atributo reoffer que es de tipo boolean asignándole el valor FALSE.

Otras de las tablas en la que se incluye información relacionada con los estudiantes son:

- Tabla: *career* (carrera). Muestra la carrera que cursan los estudiantes.
- Tabla: *course* (curso). Conserva la información relativa al curso matriculado por el estudiante.

- Tabla: *country* (país). Especifica el país de procedencia del estudiante.
- Tabla: *academic_situation* (situación académica). Describe la situación del estudiante durante el tiempo de su carrera.
- Tabla: *faculty* (facultad). Indica cuál es la facultad donde el estudiante cursa su carrera.

Los consumidores

Los consumidores de la información del Sistema son los actores de los casos de uso. Estos son:

- Los secretarios docentes de las facultades,
- Los vicedecanos docentes,
- El secretario general de la UCLV,
- Los directivos de la institución y
- El administrador de la base de datos.

Funcionalidades del sistema

El SIGENU está compuesto por varios módulos que se encargan de proveer las funcionalidades necesarias para la gestión de la información correspondiente a las distintas áreas de la actividad docente (Machado and Pupo, 2011).

Aplicación Cliente de Administración: es la aplicación que permite la inserción y actualización de los usuarios y todas las funcionalidades que deban ser ejecutadas por los administradores para monitorear el correcto funcionamiento de los procesos del sistema y su seguridad. A esta aplicación solo tendrán acceso los encargados de la administración del SIGENU.

Aplicación Cliente Secretaria: es la aplicación que constituye el elemento que esencialmente permite la inserción y actualización de toda la información que se registre en el sistema. Además, permite obtener un conjunto importante de reportes muy usados cotidianamente en el ámbito de la educación superior. Consta de los siguientes módulos: Codificadores; Matrícula; Control de estudiantes; Plan de Estudio; Evaluaciones y Reportes.

En esta aplicación podrán trabajar los actores pertenecientes al rol Secretaria en los módulos de Matrícula, Control de estudiantes, Evaluaciones y Reportes y el rol Vicedecano en el de Plan de estudio y Reportes.

Otros módulos de trabajo son: el del Secretario General, que permite la entrada y actualización de la información general de la Universidad y Facultades, así como el movimiento de estudiante entre Facultades y el Módulo Web de Recuperación de Información (Recuperador), el cual permite obtener diversos reportes con los que se puede recuperar toda la información necesaria del sistema en tiempo real, al que tienen acceso los directivos de la institución.

Aplicación Cliente Secretaria

Los casos de uso de esta aplicación son: el proceso de matrícula, el control de los datos de cada estudiante y la obtención de informes.

Proceso de matrícula

Esta opción de menú permite realizar todo el proceso de matrícula a través del cual los futuros estudiantes pasarán a ser registrados en el sistema, como estudiantes de la educación superior.

En el propio proceso de matrícula se pueden realizar las operaciones de actualizar los datos de un estudiante ya registrado como nuevo ingreso; matricular a un estudiante como nuevo ingreso; consultar reportes dinámicos sobre la matrícula y por último la cerrar el proceso de matrícula.

Dentro de este proceso existen dos conceptos importantes a tener en cuenta, los cuales son: el de *nuevo ingreso* y el de *estudiante*. Cuando una persona arriba a una secretaría y es matriculado, se registra como *nuevo ingreso* y así es considerado durante todo el proceso de matrícula. Al finalizar este proceso, se realiza la operación de *cierre de matrícula*, a través de la cual todos los nuevos ingresos pasan a ser *estudiantes* en el sistema.

Control de datos de cada estudiante

En este módulo se realizan todas las operaciones con los estudiantes del sistema, o sea, esta es la parte del sistema que permite el trabajo cotidiano de las secretarías en el curso, después de haber concluido el proceso de matrícula.

Desde esta sección del módulo se puede buscar un estudiante; listar un conjunto de estudiantes; eliminar un estudiante; distribuir estudiantes; promover estudiantes y realizar operaciones de control docente y de evaluaciones.

Reportes

En esta sección del sistema se da la posibilidad de consultar varios reportes vinculados con las SUM y el CES del centro en cuestión. Los reportes disponibles son los de: Bajas, Alumnos Ayudantes, Estudiantes Activos no Cadetes y Reportes de Año.

Una vez definidos los elementos sobre la información, las funcionalidades y consumidores involucrados en la transformación y diseminación de la información, se procede a la determinación de las dimensiones de calidad a analizar en el sistema de calidad de datos.

Para lograr esta definición se entrevistó a la administradora de la base de datos del sistema y se diseñó una encuesta para ser aplicada a los consumidores de la información.

La encuesta está dirigida a evaluar cuáles de las dimensiones que se establecen como básicas en (Crippa, 2006) y que en (Batini and Scannapieca, 2006) aparecen como las más tratadas en el uso de la metodología TDQM, son las consideradas más importantes desde la perspectiva de los usuarios del sistema.

Las dimensiones evaluadas en la encuesta son:

- Completitud.
- Exactitud.
- Actualidad.
- Fiabilidad.
- Accesibilidad.
- Consistencia.

Descripción de la encuesta

- La encuesta está compuesta por dos preguntas. En la primera de ellas se indaga por la conformidad del usuario con la información almacenada en el sistema. La segunda pregunta está referida a cómo se puede mejorar la información recopilada en el sistema.
- Cada pregunta constó de varios incisos cada uno de los cuales correspondía a una dimensión de calidad de datos, los que debían ser ordenados con valores de forma que se contestase asignando el valor cinco al aspecto que tuviera más peso para el usuario en el caso de la pregunta uno y valor cuatro en el caso de la pregunta dos.

La encuesta (ver Anexo) fue aplicada a los secretarios y los vicedecanos docentes de las facultades de Economía, Mecánica, Química y Farmacia, Psicología, Ciencias Sociales y Derecho de la UCLV. La muestra estuvo compuesta por cuatro vicedecanos docentes y seis secretarios docentes de dichas facultades.

Algoritmo de procesamiento

Para el procesamiento se siguió el siguiente procedimiento:

Para cada inciso:

Calcular el promedio de las respuestas otorgadas.

Seleccionar la dimensión a la que responde el inciso si el promedio calculado está en el rango de tres a cinco.

A continuación se muestran las tablas resumen del procesamiento de la encuesta.

Pregunta 1											
	VDD				Secretario Docente						**TOTALES (promedio)**
a	3	1	4	3	1	1	1	2	4	1	**2.33**
b	1	4	5	2	2	2	2	3	2	2	**2.78**
c	5	5	2	5	3	3	5	1	3	4	**4.00**
d	1	2	1	1	4	4	4	4	1	3	**2.78**
e	4	3	3	4	5	5	3	5	5	5	**4.67**

Tabla 13. Datos de la primera pregunta.

Pregunta 2											
VDD					Secretario Docente						**TOTALES**
a	2	1	2	1	2	3	3	3	2	2	**2.10**
b	4	3	4	3	4	2	2	2	3	4	**3.10**
c	3	2	1	4	3	1	1	4	4	3	**2.60**
d	1	4	3	2	1	4	4	1	1	1	**2.20**

Tabla 14. Datos de la segunda pregunta.

Del resultado del procesamiento, las dimensiones completitud y exactitud resultaron como las de mayor interés para los usuarios del sistema.

3.2.2 Fase de medición

En este epígrafe se definen las métricas para cada una de las dimensiones de calidad que resultaron identificadas en la fase anterior. Además de establecer qué tipo de mediciones se realizarán sobre el conjunto de datos.

Métrica para la dimensión: completitud

La métrica aplicada a esta dimensión se encuentra entre las métricas de tipo subjetivas, determinando el porcentaje de campos de datos que no tienen valores en ellos. La métrica de evaluación de calidad propuesta para esta dimensión es la siguiente.

Nombre: Grado porcentual de los valores nulos.

$$f = \frac{números\ de\ valores\ nulos}{total\ de\ valores} x100$$

Rangos (niveles de aceptabilidad) según (Covella, 2005):

Valor por debajo del 45%: satisfactorio.

Valor entre el 45% y el 65%: marginal (regular).

Valor por encima del 65%: no satisfactorio. En cuyo caso se deben aplicar herramientas y técnicas de limpieza de datos.

Interpretación: cuanto más bajo sea el porcentaje mejor.

Métrica para la dimensión: exactitud

La métrica aplicada a esta dimensión se encuentra entre las métricas de tipo subjetivas, determinando el porcentaje de valores que son corregidos en comparación con el valor real. La métrica de evaluación de calidad propuesta para esta dimensión es la siguiente:

Nombre: Grado porcentual de los valores con errores.

$$f = \frac{número\ de\ unidades\ de\ datos\ con\ errores}{total\ de\ unidades\ de\ datos}\ x100$$

Rangos (niveles de aceptabilidad) según (Covella, 2005):

Valor por debajo del 45%: satisfactorio.

Valor entre el 45% y el 70%: marginal (regular).

Valor por encima del 70%: no satisfactorio. En cuyo caso se deben aplicar técnicas y herramientas de limpieza de datos.

Interpretación: cuanto más bajo sea el porcentaje mejor.

Medición

Para efectuar la medición se utilizó una instancia de la base de datos del SIGENU, del curso 2014. El análisis se realizó sobre la tabla ***student***, la cual constituye la tabla principal, que recoge la información de la base de datos.

Para la medir "el grado porcentual de los valores nulos" se usó la herramienta *DBNulos* (Porrero, 2011), que ofrece resultados estadísticos porcentuales de ausentes por atributos (nulos, ceros y cadenas vacías), además del porcentaje total de ausentes.

La medición de completitud arrojó los siguientes valores de acuerdo a los atributos de la Tabla 15.

Atributo	% de no completitud				Observaciones
	% Nulos	% Cadenas vacías	% Ceros	% Valores ausentes	
identification	0	0	-	0	

name	0	0	-	0	
middle_name	0	0	-	0	
last_name	0	0.76	-	0.76	
native_of	0	0.008	-	0.008	
birth_date	0	-	-	0	
address	0	0	-	0	
son_count	0	-	0	0	
phone	0	76.28	-	76.28	Teléfono
email	0	99.95	-	99.95	Correo
higher_education_in_date	0.005	0	-	0.005	
university_in_date	0.005	0	-	0.005	
register_date	0.005	0	-	0.005	
scale	0.29	-	78.99	79.28	
academic_index	0.29	-	70.16	70.45	Índice académico
faculty_fk	19.7	-	-	19.7	
photo	100	0	-	100	A ninguno de los estudiantes se les almacena su foto
% total de no completitud				**40.58**	

Tabla 15. Medición de la completitud.

A partir de estos resultados se calculó el % total de no completitud.

% total de no completitud: 40.58 %

De acuerdo al % de completitud obtenido, se concluye que la dimensión Completitud en la tabla *student* se comporta a un nivel de aceptabilidad satisfactorio.

Para la aplicación de la métrica exactitud se tuvieron en cuenta los siguientes criterios que se aplican al conjunto de datos:

- El atributo carnet de identidad (identification) no está definido en la tabla ni como llave primaria, ni como atributo de valor único.
- Un estudiante puede estar registrado en la tabla con varios estados (5), pudiendo estar en el estado activo una sola vez.
- La llave primaria es un campo autogenerado (clave subrogada)

- La generalidad de los campos son de tipo *character varying*.
- Todos los estudiantes activos deben tener estar en una facultad.
- Todos los estudiantes activos deben estar en una carrera.
- La totalidad de los estudiantes registrados deben tener un valor en el campo entry_source_fk (procedencia).
- Todos los estudiantes deben tener un estado.
- Todos los estudiantes deben tener un tipo.
- Todos los estudiantes deben tener un valor en el Municipio de residencia.

En el proceso de medición se utilizó el software *DBAnalyzer* (Porrero, 2011). Este software permite obtener sobre los atributos o conjuntos de atributos diferentes valores de estadísticas que indican errores potenciales, que son valoradas con el Administrador de la base de datos quien confirma la existencia o no del error.

En el análisis realizado se combinaron algunos atributos para determinar así la exactitud de los datos.

Se combinaron los atributos de la siguiente forma:

- carnet de identidad (identification) con el estatus de los estudiantes (student_status).
- lugar de procedencia (native_of) con el tipo de estudiante (student_type).
- carnet de identidad (identification) con el name (nombre), con el primer apellido (middle_name), con el segundo apellido (last_name).

De la medición realizada usando este software se obtuvo:

Atributo	% de inexactitud	Observaciones
identification y student_status	0.1	8 registros aparecen como filas duplicadas
native_of y student_type	56.2	Existen 351 registros de estudiantes extranjeros con error en el atributo native_of.
Identification, name, middle_name y last_name	0.002	Carnet de identidad y nombre y apellidos del estudiante
birth_date, identification, student_status_fk y	62.7	Considerando los estudiantes del curso regular diurno activos. Hay 27 registros

course_type_fk		con errores en los datos de la fecha de nac.
entry_source_fk	0,05	Considerando que hay 1957 registros en que procedencia del estudiante es 'Ninguna' o 'Desconocida'
% total de inexactitud	**23.81**	

Tabla 16. Medición de la exactitud.

A partir de estos resultados se calculó el % total de inexactitud.

% total de inexactitud: 23.81%

De acuerdo al % de inexactitud obtenido, se concluye que la dimensión Exactitud en la tabla *student* se comporta a un nivel de aceptabilidad satisfactorio.

3.3 Conclusiones del capítulo

En este capítulo se aplica la metodología TDQM al sistema SIGENU de la UCLV. Fue aplicada a la base de datos operacional de dicho sistema, específicamente sobre la tabla *student* y al módulo cliente secretaria, incluyendo en este el rol de secretario docente y vice decanos docentes de las facultades de la misma universidad. Al ser aplicada la fase de definición de dicha metodología se realiza una encuesta a los usuarios antes mencionados con el objetivo de determinar cuáles eran las dimensiones de calidad con mayor peso para esos usuarios. Las dimensiones completitud y exactitud fueron las que resultaron como las más importantes a tener en cuenta.

Al ser aplicada la fase de medición, se define una métrica para la dimensión completitud, cuyo objetivo es determinar el gado porcentual de completitud en la tabla *student*, auxiliándose además de la herramienta *DBNulos.* El resultado arrojado demuestra que existe un alto grado de completitud en los datos.

De igual manera se define una métrica para la dimensión exactitud, cuyo objetivo es determinar el grado porcentual de los valores con errores en la tabla *student*, auxiliándose de la herramienta *DBAnalyzer*. El resultado obtenido demuestra que no existe un alto grado de errores en los datos.

CONCLUSIONES

Se determinaron los términos relacionados con calidad de datos, lográndose un acercamiento a sus diferentes definiciones en que todas corroboran su carácter multifacético por el amplio conjunto de dimensiones que la definen, estableciéndose que uno de los elementos al que mayor importancia se otorga es el relativo al uso del dato.

Se describieron las distintas metodologías de calidad de datos existentes en la literatura, las que fueron comparadas a partir de aspectos tales como el tipo de datos, los tipos de sistemas de información y las fases de trabajo, entre otros aspectos.

Se seleccionó y aplicó la metodología TDQM a la base de datos operacional del SIGENU de la UCLV aplicándoseles las dos primeras fases de dicha metodología. En la fase de definición se determinaron, como las dimensiones de mayor interés para el uso de los datos, las dimensiones completitud y exactitud.

En la fase de medición se especificaron las métricas para ambas dimensiones. Estas se aplicaron al conjunto de datos y de los resultados obtenidos de la evaluación, se concluye que la base de datos ofrece un aceptable nivel de calidad.

RECOMENDACIONES

La definición de un sistema de calidad de datos es un proyecto con varias fases, entre ellas, la fase de limpieza y mejora son fundamentales ya que estos proyectos no solo persiguen establecer el nivel de calidad, sino también lograr cada vez más calidad.

Es por eso que una recomendación de este trabajo es continuar, en un trabajo posterior, las fases que para la metodología TDQM se establecen en el sentido del análisis y la mejora de la calidad.

Otro elemento que se recomienda para el completamiento del estudio teórico comenzado en esta investigación es la aplicación de otra metodología a la base de datos del Sigenu desarrollando todas sus fases, de forma tal que se comparen los resultados obtenidos de la aplicación de una y otra.

Además también se recomienda completar el trabajo sobre la base de datos del mercado de datos implementado en el Sigenu, aplicando la metodología DWQ.

ANEXO

Encuesta para establecer las dimensiones de calidad de datos del SIGENU.

ENCUESTA

MODULO: CLIENTE SECRETARIA.

CONJUNTO DE DATOS: BASE DE DATOS OPERACIONAL DEL SISTEMA SIGENU EN LA UCLV.

Cargo:

Secretario docente: ___Vicedecano docente: ___

Pregunta 1.

¿Utiliza el SIGENU para hacer los informes habituales en su trabajo?

SI: ___ NO: ___

Si respondió que SI (Ordene las siguientes razones con los valores consecutivos, asignándoles valor cinco a la causa que considere de mayor peso).

a) ___ Los datos de su facultad, almacenados en el sistema, son confiables.
b) ___ Los datos están actualizados.
c) ___ Todos los datos que necesita están en el sistema.
d) ___ Los datos son obtenidos fácilmente a través del sistema.
e) ___ Los datos, al ser extraídos, están libres de contradicciones.

Si respondió que NO (Ordene las siguientes razones con los valores consecutivos, asignándoles valor cinco a la causa que considere de mayor peso).

a) ___ Los datos de su facultad, almacenados en el sistema, no son confiables.
b) ___ Los datos no están actualizados.
c) ___ No todos los datos que necesita están en el sistema.
d) ___ Los datos no son obtenidos fácilmente a través del sistema.
e) ___ Los datos, al ser extraídos, no están libres de contradicciones.

Pregunta 2.

Si se le pidieran sugerencias para mejorar la información que almacena el SIGENU, usted propondría (Ordene las siguientes razones con los valores consecutivos, asignándoles valor cuatro a la causa que considere de mayor peso).

a) ___ Agregar nuevos datos no recogidos en este momento.
b) ___ Validar mejor los datos para evitar errores en su recogida.
c) ___ Tener acceso a datos almacenados que son difíciles o imposibles de obtener para hacer informes.
d) ___ Utilizar otros valores para actualizar la información recogida.

REFERENCIAS BIBLIOGRÁFICAS

ABATE, M. L., DIEGERT, K. V. & ALLEN, H. W. 1998. A hierarchical approach to improving data quality. *Data Quality Journal,* 4**,** 365-369.

BATINI, C., CAPPIELLO, C., FRANCALANCI, C. & MAURINO, A. 2009. Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR),* 41**,** 16.

BATINI, C. & SCANNAPIECA, M. 2006. Methodologies for Data Quality Measurement and Improvement. *Data Quality: Concepts, Methodologies and Techniques***,** 161-200.

BOTTURA FILHO, J. A. 2007. Programa de Pós-Graduação Latu Sensu MBIS–Executivo em Ciência da Computação.

BRACKSTONE, G. 1999. Managing data quality in a statistical agency. *Survey methodology,* 25**,** 139-150.

BURNS, E. M., MACDONALD, O. & CHAMPANERI, A. Data quality assesment methodology: A framework. Joint Statistical Meetings-Section on Government Statistics, 2000. 334-337.

CALAZANS, A. T. S. 2008. Information quality: concepts and applications. *Transinformação,* 20**,** 29-45.

CARLO, B., DANIELE, B., FEDERICO, C. & SIMONE, G. 2011a. A data quality methodology for heterogeneous data. *International Journal of Database Management Systems (IJDMS),* 3.

CARLO, B., DANIELE, B., FEDERICO, C. & SIMONE, G. 2011b. A data quality methodology for heterogeneous data. *International Journal of Database Management Systems,* 3**,** 60-79.

CARO, A., FUENTES, A. & SOTO, M. A. 2013. Desarrollando sistemas de información centrados en la calidad de datos. *Ingeniare. Revista chilena de ingeniería,* 21**,** 54-69.

COVELLA, G. J. 2005. *Medición y evaluación de calidad en uso de aplicaciones web.* Facultad de Informática.

CRIPPA, G. 2006. Una metodologia per la valutazione dei costi della non qualità dei dati.

CYKANA, P., PAUL, A. & STERN, M. DoD Guidelines on Data Quality Management. IQ, 1996. 154-171.

DAZA, A., DE LA TORRE, P., ZEPEDA, V. V. & VILLEGAS, C. M. 2012. Hacia un Modelo de Madurez para la Gestión de Calidad de Datos en Inteligencia de Negocios.

DEJAEGER, K., HAMERS, B., POELMANS, J. & BAESENS, B. A novel approach to the evaluation and improvement of data quality in the financial sector. Proceedings of the 15th International Conference on Information Quality, 2010.

DOMINGO, G., BUCCELLA, A. & CECHICH, A. Un marco de trabajo para analizar y mejorar la calidad de datos dentro de su ciclo de vida. XIII Congreso Argentino de Ciencias de la Computación, 2007.

EMRAN, N. A., ABDULLAH, N. & MUSTAFA, N. 2012. A Review of Failure Handling Mechanisms for Data Quality Measures.

EPPLER, M. J. & MUENZENMAYER, P. Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology. IQ, 2002. 187-196.

FALORSI, P., PALLARA, S., PAVONE, A., ALESSANDRONI, A., MASSELLA, E. & SCANNAPIECO, M. Improving the quality of toponymic data in the italian public administration. Proceedings of the ICDT, 2003.

FISHER, C. W., LAURIA, E. J. & MATHEUS, C. C. 2009. An accuracy metric: Percentages, randomness, and probabilities. *Journal of Data and Information Quality (JDIQ),* 1**,** 16.

FISHER, T. & MARINOS, G. 2003. Better decisions through better data quality management. PricewaterhouseCoopers.

GALLO, B. B. & CORRADO, M. C. V. 2009. FACULTAD DE INGENIERÍA UNIVERSIDAD DE LA REPÚBLICA Un caso de estudio en Calidad de Datos para Ingeniería de Software Empírica.

GAMA, C. O. H., JAIMES, B. M. & PAZ, L. P. 2013. Método de madurez para la calidad de los datos. *Punto de Vista,* 3.

GEIGER, J. 2004. Data quality management: the most critical initiative you can implement. *Data Warehousing, Management and Quality, Paper***,** 098-29.

GUZMÁN, E. C. & DE LLERGO, L. M. L. 2013. MODELO ESTRATÉGICO PARA LA ADMINISTRACIÓN DE LA CALIDAD DE LA INFORMACIÓN FINANCIERA.

JAASKELAINEN, O., KROPSU-VEHKAPERA, H. & HAAPASALO, H. 2011. LIFECYCLE VIEW ON PRODUCT DATA QUALITY DIMENSIONS.

JEUSFELD, M. A., QUIX, C. & JARKE, M. 1998. Design and analysis of quality information for data warehouses. *Conceptual Modeling–ER'98.* Springer.

KOVAC, R., LEE, Y. W. & PIPINO, L. Total Data Quality Management: The Case of IRI. IQ, 1997. 63-79.

LEE, Y. W., STRONG, D. M., KAHN, B. K. & WANG, R. Y. 2002. AIMQ: a methodology for information quality assessment. *Information & management,* 40**,** 133-146.

LIDIANSA, M. 2014. *Developing Data Quality Metrics for a Product Master Data Model.* TU Delft, Delft University of Technology.

LONG, J. A. & SEKO, C. E. 2005. A cyclic-hierarchical method for database data-quality evaluation and improvement. *Advances in Management Information Systems-Information Quality (AMIS-IQ) Monograph*.

LOSHIN, D. 2001. *Enterprise knowledge management: The data quality approach*, Morgan Kaufmann.

LOSHIN, D. 2006. Monitoring Data Quality Performance Using Data Quality Metrics: A White Paper. *Informatica. November*.

LUEBBERS, D., GRIMMER, U. & JARKE, M. Systematic development of data mining-based data quality tools. Proceedings of the 29th international conference on Very large data bases-Volume 29, 2003. VLDB Endowment, 548-559.

MACHADO, R. L. & PUPO, O. G. R. 2011. Módulo para el control de la baja estudiantil en el SIGENU. *Ciencias Holguín,* 17.

MILANO, D., SCANNAPIECO, M. & CATARCI, T. 2006. Design and implementation of a peer-to-peer data quality broker. *Interoperability of Enterprise Software and Applications.* Springer.

MOGES, H.-T., DEJAEGER, K., LEMAHIEU, W. & BAESENS, B. 2013. A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Information & Management,* 50**,** 43-58.

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

PIPINO, L. L., LEE, Y. W. & WANG, R. Y. 2002. Data quality assessment. *Communications of the ACM,* 45**,** 211-218.

PORRERO, B. L. 2011. *Limpieza de datos: Reemplazo de valores ausentes y estandarización.* Doctoral.

REDMAN, T. C. 1998. The impact of poor data quality on the typical enterprise. *Communications of the ACM,* 41**,** 79-82.

REDMAN, T. C. 2001. *Data quality: the field guide*, Digital press.

REYES, Y. M. 2012. Espacio de comunicación e intercambio para la Comunidad Técnica Cubana de PostgreSQL. *Serie Científica,* 5.

ROSALES, J. A. G. & HERRERA, C. F. U. 2012. Modelos de Madurez en los Datos de una Organización; Caso de Estudio Universidad Católica Boliviana "San Pablo" Cochabamba. *ACTA NOVA,* 5.

SCANNAPIECO, M., VIRGILLITO, A., MARCHETTI, C., MECELLA, M. & BALDONI, R. 2004. The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems. *Information systems,* 29**,** 551-582.

SEDÓ, M. Z. 2007. Calidad de Datos.

SHANKARANARAYANAN, G. & WANG, R. IPMAP: Current state and perspectives. Proceedings of the 12th International Conference on Information Quality, 2007.

SIVOGOLOVKO, E. 2011. Evaluation of impact of data quality on clustering with syntactic cluster validity methods. Technical report, Christian-Albrechts University.

STRONG, D. M., LEE, Y. W. & WANG, R. Y. 1997. Data quality in context. *Communications of the ACM,* 40**,** 103-110.

SU, Y. & JIN, Z. 2006. A methodology for information quality assessment in the designing and manufacturing process of mechanical products. *Information Quality Management: Theory and Applications***,** 190-220.

TAYI, G. K. & BALLOU, D. P. 1998. Examining data quality. *Communications of the ACM,* 41**,** 54-57.

URIBE, I. A. 2010. GUÍA METODOLÓGICA PARA LA SELECCIÓN DE TÉCNICAS DE DEPURACIÓN DE DATOS.

VALVERDE, C., MAROTTA, A. & VALLESPIR, D. 2009. Análisis de la Calidad de Datos en Experimentos en Ingenierıa de Software.

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

WANG, R. Y. & STRONG, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems***,** 5-33.

WIJNHOVEN, F., BOELENS, R., MIDDEL, R. & LOUISSEN, K. 2007. Total data quality management: A study of bridging rigor and relevance.