CUORE

enero 2010 • nº 42 • 6,50 €

Círculo de Usuarios Oracle de España

XIX Congreso CUORE

NEC

2010: el PC de sobremesa ya es historia en el mundo empresarial

ORACLE

Reducción de costes de TI en el entorno económico actual

Exprimiendo el valor de los datos con Oracle Spatial

SUN MICROSYSTEMS

Oracle Exadata V2

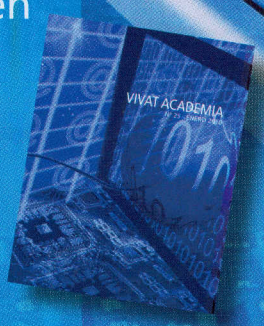
NEW PATH

Gestión de recursos de CPU en entornos consolidados usando DBRM

LOPDGEST

¿Cumplimos la normativa en materia de protección de datos?

VIVAT ACADEMIA



Sumario

Edita: Cuore. www.cuore.es
Editor: Rafael Rojo
Consejo Editorial: José Manuel López, Máximo Aborruza, Rafael Rojo, José Manuel Peláez, Marta Eguskiza y David Abreu.
Producción gráfica: Moncaba
Depósito Legal: M-24195-1992

CUORE

Estimado compañero/a:

Te presentamos este nuevo número de nuestra revista, confeccionado, como siempre, con especial cariño y cuidado, con el ánimo y la intención de que nos sientas muy próximos a ti.

En el ejemplar que tienes en tus manos vas a encontrar, en primer lugar, un interesante artículo de Cristina Mateos, de NEC Computers, sobre cómo el auge del "cloud computing" va a transformar el panorama físico y estético de nuestras oficinas porque va a favorecer la desaparición de los PCs de sobremesa tradicionales y su sustitución por equipos más "ligeros" donde la conexión a la red y el escritorio virtualizado serán los elementos principales. Precisamente en estas mismas fechas, Oracle está organizando una serie de eventos encaminados a explicar cómo aprovechar en las empresas el modelo "cloud computing", cómo funciona esta tecnología y por qué conviene usarla. Desde aquí os animamos a asistir a estas interesantes sesiones.

Javier Cabrerizo de ORACLE, aborda el tema de cómo reducir costes en Tecnologías de la Información en tiempo de crisis. Señala Javier en su artículo la gestión de los datos y optimización del almacenamiento así como la virtualización de recursos como aspectos muy a tener en cuenta a la hora de optimizar los costes de TI.

Hemos incluido también, un extenso artículo de Alfredo García de SUN Microsystems, en el que se hace un exhaustivo y pormenorizado análisis de Exadata V2, la "Sun Oracle Database Machine". Gran apuesta de SUN y ORACLE para revolucionar el rendimiento en el tratamiento de la información, representa sin lugar a dudas el futuro ya presente. La lectura del artículo es altamente recomendable.

Puedes leer también un artículo de Arturo Gutiérrez de Newpath sobre el gestor de recursos de la base de datos (DBRM), ese gran desconocido para la gran mayoría, a pesar de que fue implementado en el núcleo de la base datos ya en la versión 8i. Arturo, con su capacidad didáctica y su estilo ameno, hace un repaso de la evolución de este gestor hasta la actual 11g R2 y explica cómo usarlo para gestionar eficazmente los recursos de CPU.

Juan A. Espin de ORACLE y Mariana Estrada de LOPDGEST, incluyen sendos artículos sobre Oracle Spatial y normativa LOPD, respectivamente. Por supuesto, incluimos nuestra tradicional separata Vivat Academia que incluye un jugoso artículo de Ismael Caballero, Coral Calero y Mario Piattini, en el que nos plantean un modelo de información para medir la calidad de los datos.

Sólo me resta desearte que disfrutes de esta edición de nuestra revista e indicarte que puedes participar en la realización de la misma, escribiendo y haciéndonos llegar cualquier artículo que consideres pueda resultar de interés a nuestros compañeros asociados.

Por último, te invito a que elabores alguna ponencia para nuestro próximo Congreso que por acuerdo de la Junta Directiva tendrá lugar en Madrid el próximo mes de Octubre. Recibe un fuerte abrazo y espero verte, si no nos vemos antes, en el Congreso de Madrid.

José Manuel López
Presidente de Cuore

VIVAT ACADEMIA

Nº 25 - ENERO 2010

DQMIM: UN MODELO DE INFORMACIÓN PARA LA MEDICIÓN DE CALIDAD DE DATOS

Ismael Caballero, Coral Calero, Mario Piattini
Universidad de Castilla–La Mancha
Grupo Alarcos - Instituto de Tecnologías y Sistemas de la Información
Pº de la Universidad 4, 13071 Ciudad Real (Spain)
{Ismael.Caballero, Coral.Calero, Mario.Piattini}@uclm.es

1. RESUMEN

La medición es una actividad clave en la gestión de la Calidad de los Datos (DQ). Dada la naturaleza específica de los datos, establecer medidas acerca de su grado de calidad resulta en ocasiones una labor tremendamente ardua. En la bibliografía existente se pueden encontrar varios trabajos que realizan aportaciones a este campo. Aunque estén referidas a los mismos conceptos, dichas propuestas están generalmente circunscritas a problemas específicos, con lo que generalizar para reutilizar en distintos contextos resulta prácticamente imposible. Además, cada uno de los autores utiliza su propia terminología, motivando una falta de unificación en la nomenclatura, lo que aún más dificulta la reutilización. El objetivo principal de este artículo es proponer una nomenclatura común a través de un Modelo de Información de Medición de Calidad de Información (DQ-MIM) basándose en el estándar ISO/IEC 15939. Para ello se analizan a través de una serie de preguntas guía, cómo los distintos autores han tratado los aspectos más relevantes de la medición de calidad de datos. La nueva terminología, permitirá describir mejor los Planes de Medición de Calidad de Datos que se hagan en ciertos contextos. Con el objetivo de hacer operativo el modelo de información presentado, se ha diseñado un esquema XML para dar soporte a la documentación de dichos planes de medición.

Palabras clave: Calidad de los datos de medición, ISO/IEC 15939, Modelo de Información de Medición de Calidad de Datos

2. INTRODUCCIÓN

Normalmente, una organización se da cuenta de la importancia de la calidad de los datos (en adelante, DQ) cuando los problemas debido a una carencia de ella acaban afectado negativamente al rendimiento del negocio. Una vez que esto ha ocurrido, los ejecutivos necesitan cuantificar el impacto real de estos problemas a diferentes niveles (de organización, económicos, de satisfacción del cliente, o satisfacción de los empleados) para poder clasificarlos y esbozar planes de mejora de DQ que los mitiguen [22].

Cualquier plan de mejora de DQ debe comenzar con la evaluación de los escenarios afectados para identificar las raíces comunes de los problemas detectados, y determinar si el alcance es local al escenario o global a todo el departamento o a toda la organización. Esta evaluación implica tener valores de las medidas de DQ de los recursos de datos organizacionales implicados. La intención principal de estas medidas es proporcionar un sentido cuantitativo sobre la cantidad de calidad de datos (expresadas a través de las dimensiones de la calidad) que se tiene, a fin de habilitar las correspondientes acciones para una gestión adecuada [13]. Aunque en la bibliografía existen varias propuestas sobre medición de DQ, no es un problema que esté satisfactoriamente resuelto y existen todavía muchos retos abiertos de investigación en el área [3]. Creemos que uno de estos retos consiste en la unificación de los diferentes términos proporcionados por los diferentes autores para los mismos conceptos.

Con el fin de lograr este objetivo, se partió de la idea de alinearlos en torno a un estándar internacional sobre medición; dado que no existe ninguna propuesta para datos, se pensó que algún estándar de medición de software podría ser un buen punto de partida, pues sería conocido por muchas organizaciones, y se facilitaría la labor de transferencia de los resultados de investigación. De entre todos hemos seleccionado ISO / IEC 15939 [21] para esta propuesta, pues define un modelo de información de medición (MIM), que será la base del Modelo de Información para la Medición de la Calidad de Datos (DQMIM) que se describe en la sección 2. Nos gustaría resaltar que el objetivo de este trabajo no es en sí desarrollar medidas de calidad de datos, sino acercar al lector los conceptos más relevantes y proporcionar una nomenclatura común de dichos conceptos para facilitar el proceso de definición de medidas.

Teniendo en cuenta la definición de calidad de datos ("fitness for use"), tiene sentido pensar que la medición de calidad de datos depende fuertemente del contexto en que se usen los datos. Así una persona, dependiendo del rol que esté desempeñando en un determinado proceso de negocio, tendrá una percepción distinta del nivel de calidad de datos de un recurso de datos organizacional. Así, si varios roles están evaluando la calidad de datos de los mismos datos pero desde diferentes puntos de vista, será necesario definir distintas medidas incluso para las mismas características o dimensiones de calidad. En el siguiente apartado se introducen los aspectos que hay que tener en cuenta para definir medidas de calidad de datos. Dada la complejidad del proceso, es conveniente hablar de la necesidad de definir **Planes de Medición de Calidad de Datos**. En este sentido, la aportación de este trabajo es presentar una nomenclatura que sirva de base para identificar los elementos que intervienen en los planes de medición.

3. UN MODELO DE CALIDAD DE LOS DATOS DE MEDICIÓN DE INFORMACIÓN (DQMIM)

3.1 INCLUSIÓN DE TRABAJO EN LA REVISIÓN DE LA BIBLIOGRAFÍA RELACIONADA

Si bien se han encontrados varios trabajos en la bibliografía, hemos elegido aquellos presentados en [11, 22, 24, 26, 34] tras haber realizado una revisión sistemática, pues son los que pueden ser considerados como los más representativos. No obstante, y cuando así fuese necesario, hemos considerado para la elaboración de este trabajo otros que pueden tener cierta relevancia. Por ejemplo, aunque se ha supuesto que [22] reúne y reemplaza los trabajos previos realizados por los investigadores del MIT (considerados como los padres del área), otros trabajos como los propuestos en [31, 41] han sido estudiados y referenciados

A la hora de llevar a cabo nuestra investigación para la identificación de los elementos que intervienen en el proceso de definición de un plan de medición de calidad de datos, pensamos que una buena estrategia podría consistir en realizar un análisis basado en preguntas del tipo "¿Quién?/¿Cómo?/¿Por qué? / ... para realizar la comparativa entre las distintas propuestas de medición seleccionadas y alinearlos con los términos equivalentes propuestos en ISO/IEC 15939. La Tabla 1 recoge por

columnas las preguntas, los términos proporcionados por ISO/IEC 15939, y en una tercera columna se muestran los términos equivalentes encontrados en la bibliografía. Con el fin de alinear DQMIM al estándar, se usaron los mismos nombres que éste proporcionaba.

3.2 ¿POR QUÉ MEDIR?

Aunque pudiera parecer demasiado obvio, parece interesante que nuestra primera tarea en la definición de este DQMIM debe ser revisar el significado del concepto de "**medida**". De acuerdo con el estándar ISO / IEC 15939, "**medir**" es "*hacer una serie de operaciones con el objeto de determinar el valor de una representación cuantitativa o categórica de uno o más atributos*"; además "*las medidas deben tener un propósito claramente definido*". Este objetivo para la medición de la calidad de datos de un escenario es satisfacer una determinada "**necesidad de información**" para gestionar objetivos, metas, riesgos y problemas (véase Tabla 1). De acuerdo al estándar, el término "*métrica*" no debería seguir siendo utilizado como sinónimo de "**medida**".

Conociendo una "**necesidad de información**" determinada, definir un plan de medición se orienta a determinar:

- qué medir,
- dónde están los objetos que se van a medir y de quién son,
- la forma de medir estos objetos,
- cuántos objetos son necesarios inspeccionar a fin de tener una evidencia estadísticamente significativa sobre el grado de satisfacción de las necesidades de información,
- quién debe diseñar e implementar los procedimientos de medición,
- quién puede realizar la medición (teniendo en cuenta que debería tener acceso a los recursos de datos, y que si no los tiene, los resultados de las medidas serán diferentes),
- a quién deben ser entregados los resultados de la medición y, finalmente,

Pregunta	Related terms by ISO/IEC 15939	Related term in DQ Literature
¿Qué?	Concepto Medible	Dimensiones de Calidad de Datos [11, 22, 24], Criterios de Calidad de Información [26]
¿Quién?	Implicado (Usuario de la Medida, Analista de Medición, Propietario del proceso de Medición, Proveedor de Datos, Usuario de los datos)	Clientes de Datos [24], Productores de Datos, Custodiadores de Datos, Consumidores y Gestores de Datos [42], Personas que crean o actualizan un grupo de datos [11]
¿Cómo?	Medida (Medida base, Medida Derivada, Indicador), Forma de Medir (Método de Medición, Función de Cálculo de Medición, Criterio de Decisión)	Medición [24], Juicios o Valoración de Criterios de Calidad [11], Puntuación de Métodos de Valoración [26]
¿Cuántos implicados?	-	Número de usuarios que proporciona información subjetiva sobre la calidad de un dato [17]

Tabla 1. Análisis guiado pro preguntas para la creación de una nomenclatura común de los términos de medición de DQ encontrados en la Bibliografía.

Concepto	Significado en ISO/IEC 15939	Término en el área DQ
Necesidad de Información	Una abstracción necesaria para gestionar objetivos, metas, riesgos y problemas.	Objetivo de valoración de Calidad de Información [11], Problema [22], Proyectos fundamentales [14]
Medida	Una representación cuantitativa o categórica de uno o más atributos.	

Tabla 2. Conceptos proporcionados por ISO/IEC 15939 para responder a "¿por qué medir?"

- cuando debe realizarse la medición a fin de no interferir en el propio proceso de medición, o incluso en el proceso de fabricación de información.

En el campo de DQ, hay algunas equivalencias para el término "**necesidad de información**": por ejemplo, en [11] se establece el término "*Objetivos de valoración de calidad de información*". Algunos ejemplos de estas necesidades de información proporcionados por este autor son "*comprender el estado de la calidad en una base de datos*", "*identificar la información de fabricación que requieren los procesos de mejora*", o "*evaluar una determinada dimensión de calidad de datos*", o incluso como [27] propone, "*medir los problemas de una base de datos relacional*". También en [14] se identifican lo que se da en llamar "*proyectos fundamentales de calidad de datos*" que bien podría ser asimilado como sinónimos del concepto de "*necesidad de información*".

En nuestro ejemplo, los editores quieren medir la fiabilidad y compleción de las noticias, porque se ha demostrado que estos dos factores son muy importantes para su negocio (por ejemplo, una encuesta de opinión de calidad a sus lectores ha reflejado estos dos factores como los más importantes, y que les condicionarían los suficiente como para volver a visitar la web del periódico on-line). Por lo tanto, su necesidad de información se describen claramente lo siguiente: "*Los editores quieren saber si la calidad de los datos de las noticias publicadas en su web satisface las expectativas de sus clientes*".

3.2 ¿QUÉ Y DÓNDE MEDIR?

Una persona responsable de mejorar la calidad de los datos debería analizar la Especificación de Requisitos de

Concepto	Significado en ISO/IEC 15939	Término en el área DQ
Concepto Medible	Un concepto cuya medición satisface una determinada necesidad de información.	Dimensiones de Calidad de Datos [11, 22, 24], Criterios de Calidad de Información [26]
Almacén de Datos	Una colección organizada y persistente de datos que permite su recuperación.	Base de datos relacional, Base de Datos Objeto Relacional, Base de datos XML, Hoja de cálculo.

Tabla 3. Conceptos de ISO/IEC 15939 usados para modelar las preguntas "¿qué?" y "¿dónde?"

usuario de Calidad de Datos (DQ-URS) en busca de lo que se requiere medir y donde están los objetos sobre los que hay que medir.

La respuesta a la pregunta "**qué**" son los "**conceptos medibles**" para los "**atributos medibles**" de las "**entidades**" que el usuario considera que está implicada en el proceso de medición. Los conceptos medibles son lo que en el campo DQ han sido tradicionalmente llamados como **dimensiones de calidad de datos**. Tal vez, éste es uno de los conceptos más tratados en la bibliografía porque es la base para entender lo que significa DQ para los distintos tipos de usuarios y de contextos [33]. Se pueden encontrar descripciones de dimensiones de calidad de datos y discusiones sobre cuáles son las dimensiones más importantes para un determinado contexto en [1, 3, 11, 12, 16, 22] por citar algunos. Aunque muchos investigadores han identificado aquellas dimensiones que mejor se adaptan a su problema, y que recientemente fue publicado el estándar ISO/IEC 25012 [19] (que reúne las dimensiones de calidad de datos para sistemas de información-véase Tabla 4), puede decirse que todavía no existe un conjunto universal de dimensiones (conceptos medibles) de DQ que pueda ser considerado válido para cualquier contexto, así como tam-

Inherentes Dimensión	Descripción
Exactitud (*Accuracy*)	El grado en el cual el dato tiene atributos que correctamente representan el valor correcto del atributo intencionado de un concepto o evento en un contexto específico de empleo.
Consistencia (*Consistency*)	El grado en el cual el dato tiene los atributos que son libres de contradicción y son coherente con otros datos en un contexto específico de uso.
Actualidad (*Currentness*)	El grado en el cual el dato tiene los atributos que son del período correcto en un contexto específico de uso.
Inherentes y Dependientes del sistema	
Accesibilidad (*Accessibility*)	El grado en el cual el dato puede ser accesado en un contexto específico de uso, en particular por la gente que necesita el soporte de tecnología o una configuración especial debido a alguna inhabilidad (incapacidad).
Confidencialidad (*Confidentiality*)	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.
Precisión (*Precision*)	El grado en el cual el dato tiene atributos que son exactos o que proporcionan la discriminación en un contexto específico de uso.
Entendibilidad (*Understandability*)	El grado en el cual el dato tiene atributos que le permiten ser leído e interpretado por usuarios, y es expresado en lenguajes apropiados, símbolos y unidades en un contexto específico de uso.
Dependientes del Sistema	
Disponibilidad (*Availability*)	El grado en el cual el dato tiene atributos que le permiten ser recuperados por usuarios autorizados y/o aplicaciones en un contexto específico de uso.
Recuperabilidad (*Recoverability*)	El grado en el cual el dato tiene atributos que le permiten mantener y conservar un nivel especificado de operaciones y calidad, aún en caso de falla, en un contexto específico de uso.

Tabla 4. Dimensiones de Calidad (conceptos medibles) de Datos según ISO/IEC 25012 [19]

poco existe un conjunto exhaustivo de medidas para estas dimensiones [9].

De todos modos, la clasificación de las dimensiones de DQ propuesta por Strong et al. en [36], es considerada como punto de partida para la mayoría de los trabajos de investigación y para la mayoría de las iniciativas de gestión de DQ. Estas dimensiones son agrupadas en cuatro categorías, tal y como se representa en la Tabla 5.

Finalmente, se puede decir que la relación existente entre "**necesidades de información**" y "**concepto medible**" podría establecerse como que "*una necesidad de información podría ser satisfecha mediante la combinación de uno o más conceptos medibles*".

En nuestro ejemplo de las noticias, los conceptos medibles son "**fiabilidad**" y la "**compleción**". Atendiendo a las necesidades de los editores, asumimos que entienden por "*fiabilidad*" como la "*el grado en que las noticias que llegan provienen de una fuente fiable*", mientras que por "**compleción**" entienden "*el grado en que las noticias tienen datos para todos los campos identificados*"

La respuesta a la pregunta "**dónde**" es en "*las entidades que tienen atributos medibles*". Los posibles atributos medibles puede ser cualquiera de los identificados por [24]. A fin de ganar en generalidad, se va a usar el término "**almacén de datos**" (según el estándar ISO/IEC 15939) para referirse a cualquier "**categoría de las entidades**"

Categoría	Conceptos Medibles de Calidad de Datos
Intrínsecas	Exactitud, Objetividad, Credibilidad, Reputación
Contextual	Relevancia, Valor Añadido, Oportunidad, Compleción, Cantidad de Datos.

Tabla 5. Dimensiones de Calidad(conceptos medibles) de Datos según [36]

dedicados a almacenar o presentar datos (por ejemplo, cualquier base de datos relacional, o cualquier documento XML). En [11] también se incluyen a los archivos o procesos que deben medirse. Incluso en [6] se consideran ficheros RDF de Web Semántica. No obstante, aunque se asume que las entidades se refieren principalmente a almacenes de datos, hay otros elementos que son susceptibles de ser medidos. Los párrafos siguientes ofrecen un análisis acerca de los atributos medibles identificados por [24] que pueden también ser medidos. Dependiendo del tipo de atributo medible a medir, algunas entidades son susceptibles de ser asignadas medidas clasificadas como estructurales (e.g. modelos de datos, presentación de los datos y las políticas de calidad de datos), mientras que en otros casos se puede hablar de medidas relacionadas con su contenido (valores de datos y campos de datos) [13].

Para el caso de los **modelos de datos**, se puede hacerse una distinción entre los modelos conceptuales y modelos lógicos. Para más información sobre cómo medir los modelos conceptuales, se puede encontrar una descripción de los correspondientes conceptos medibles en los trabajos [2, 35], mientras que las correspondientes medidas están cubiertos en [15, 23, 25, 30]. Para el modelo lógico, algunas medidas propuestas que pueden ser particularmente útiles, se pueden encontrar en [8, 29]. Queremos resaltar que aunque los trabajos anteriores podrían no tratar directamente la calidad de los valores de datos, los resultados mostrados se pueden utilizar como base para definir medidas de calidad de datos (típicamente de tipo densidad que se calculan como un ratio), y por esa razón consideramos interesante incluirlos en este análisis.

Por otro lado, las mediciones de la calidad de los datos para los valores de los datos están destinadas a obtener valores sobre los datos contenidos en los almacenes de datos. Como éste ha sido uno de los principales focos de la bibliografía de DQ, ha sido ampliamente estudiado a través de los distintos trabajos identificados en la sección 2.1.

El valor de un dato se toma siempre de un **dominio del**

dato. Si los dominios no están correctamente definidos, los almacenes de datos se pueden llenar de datos incorrectos. Dado que un dominio de datos es en sí mismo un conjunto de datos, es posible medir la calidad de datos a través de conceptos medibles, como la integridad o la precisión de un dominio [34].

Otra categoría importante de entidades que merece la pena incluir en este análisis son aquellas dedicadas a presentar los datos a los usuarios: las interfaces de usuario. La importancia de este tipo de entidad radica en el hecho de que son el principal contacto del usuario con los datos y la forma en que los datos son mostrados a / recolectados de los usuarios. Por esa razón, las interfaces de usuario también puede ser medidas desde el punto de vista de la calidad de los datos (categoría representacional -véase Tabla 5-). Por ejemplo, los trabajos presentados en [10] están orientados a medir la calidad de los datos representacional de los portales web como interfaces de usuario que son.

Está ampliamente aceptado que dado el carácter subjetivo de la calidad de datos, es necesario establecer una serie de reglas de negocio que describan cuándo un dato puede o no tener un nivel de calidad aceptable para su uso en la organización. Típicamente estas reglas de negocio se expresan mediante políticas organizacionales de DQ [22]. Estas políticas organizacionales de DQ son una manera de "*universalizar*" las lecciones aprendidas a través de las diferentes experiencias sobre cómo gestionar los conceptos medibles, los riesgos de DQ, y cómo modificar los modelos de datos y de procesos para adaptarla a las "*mejores prácticas DQ*". En este sentido, en [24] se propone también medir la calidad de las políticas organizacionales desde el punto de vista de DQ.

La relación existente entre los "*conceptos medibles*" y "*entidades*" es que "*un concepto medible puede incluir uno o más atributos medibles que pertenecen a una entidad*".

Para el ejemplo que se está presentando en este trabajo, la entidad es el documento RSS (XML), que almacena las noticias, mientras que los atributos de medición son los valores de datos, que se pueden encontrar en los elementos con sus correspondientes atributos ("*elementos*" y "*atributos*" son definidos aquí como parte de un archivo XML [37].

3.3 ¿QUIÉN DEBE MEDIR Y DE QUIÉN SON LAS ENTIDADES QUE VAN A MEDIRSE?

A fin de evaluar y mejorar la DQ en una organización, es necesario que un equipo de trabajadores con los conocimientos y responsabilidades suficientes tanto en los aspectos de producción de información como de gestión de calidad sea capaz de identificar todos los **"actores"** implicados en el proceso de medición. Las funciones de estos actores dependen de su rol sobre los datos. [41] identifica como posibles funciones para los actores: "*Proveedores de información*", "*Productores de Información*", "*Consumidores de Información*", y "*Gestores de Información*" (aquí, el autor utiliza indistintamente los términos "*datos*" e "*información*"). Corresponde a los gestores de calidad de información diseñar, dirigir y obtener conclusiones acerca de los resultados del proceso de medición. [24] recomienda identificar a los consumidores de datos como una pieza clave del plan de medición para realizar exitosamente una medición. Es importante resaltar el hecho de que no todos los consumidores de datos tienen porqué ser necesariamente seres humanos, sino que pueden ser otros procesos de trabajo que se ejecutan sobre el mismo o diferentes sistemas de información. [34] propone no sólo identificar los consumidores de datos (o clientes), sino también describir cómo usan los datos necesarios para determinar las características y los niveles requeridos de calidad. Los mismos conceptos medibles podría medirse de diferentes maneras dependiendo del rol del actor: dos roles diferentes pueden requerir los mismos conceptos medibles, pero para satisfacer necesidades de información diferentes.

Al diseñar el plan de medición, el equipo de medición de DQ debe tener en cuenta a quién pertenecen las entidades que conteniendo datos deben ser medidas. Este tipo de actores se denomina "**propietarios de los datos**". Este

Concepto	Significado en ISO/IEC 15939	Término en el área DQ
Implicados / actores	Un individuo o una organización que patrocina las medidas y proporciona los datos o es un usuario de los resultados de las medidas	Trabajadores que crean o actualizan un grupo de datos [11]; propietarios / colectores, custodiadores o consumidores de datos [22]; Clientes de datos, gestores o productores de datos; proporcionadores de datos [42];
Almacén de Datos	Una colección organizada y persistente de datos que permite su recuperación.	Base de datos relacional, Base de Datos Objeto Relacional, Base de datos XML, Hoja de cálculo.

Tabla 6. Conceptos de ISO/IEC 15939 usados para modelar las respuestas a las preguntas quién y de quién?

hecho es importante porque, a veces, sólo pueden medir la calidad de los datos sus propietario(s) o aquellas personas a las que se ha concedido el acceso a ellos.

Para el ejemplo de las noticias, el "*proveedor de datos*" es el proveedor de noticias, los "*consumidores de datos*" son los clientes o lectores del periódico electrónico; el "*productor de datos*" es el conjunto de procesos que recibe la noticia de su proveedor de noticias y le da el formato adecuado para mostrarlos a los consumidores de datos, y por último los "*gestores de datos*" y los "*propietarios de los datos*" son los editores.

3.4 ¿CÓMO MEDIR Y CUÁNTOS DATOS ESTÁN INVOLUCRADOS EN LAS MEDICIONES?

Esta es probablemente la cuestión que requiere más atención que cualquiera de las otras tratadas a través de este trabajo. Una vez que uno o más conceptos medibles para cada necesidad de información han sido identificados a partir de la especificación de requisitos de calidad de datos de los usuarios y, estando claro cuáles son los atributos medibles que pertenecen a las entidades correspondientes, el siguiente paso es definir las medidas en sí. ISO/IEC 15939 clasifica las medidas de la siguiente manera: "**medida base**", "**medida derivada**" e "**indicadores**" (véase Tabla 6). La forma en la que un concepto medible se mide es descrita por un método de medición definido sobre los atributos medibles. El estándar identifica dos tipos de métodos de medición: objetivos y subjetivos. En el campo de DQ, esta diferencia también ha sido observada por varios autores como en [32].

Debido al carácter subjetivo de la calidad de los datos, es importante destacar la diferencia entre los conceptos de **medición** ("*medición es el acto de asignar un número a un atributo de un objeto observado*") frente a la **evaluación** ("*la clasificación de alguien o algo con respecto a su valor*"). Considerando que el primer término se destina a definir y operar con valores cuantitativos (los posible "*tipos de escalas*" son ratio e intervalo), la última tiene por objeto definir y gestionar los valores cualitativos (los tipos posibles de escala son principalmente ordinales).

Para cada medida se debe proporcionar tanto una **escala** (lo que implica seleccionar un dominio de posibles valores) como una **unidad de medida**. La Tabla 8 muestra la particularización de estos términos para el ejemplo que se está desarrollando a través del trabajo.

Dado que la calidad de datos es un concepto subjetivo, y que el tratamiento de su subjetividad ha sido ya abordado por la bibliografía, merece la pena analizar un poco este

Concepto	Significado en ISO/IEC 15939	Término en el área DQ
Medida Base	Un atributo y el método para cuantificarlos	**Métrica, Medida**
Criterio de Decisión	Umbrales numéricos u objetivos usados para determinar la necesidad de una acción, para realizar alguna investigación, o para describir el nivel de confianza en un resultado dado.	Indicador
Indicador	Una estimación o evaluación de atributos específicos derivados de un modelo con respecto a una necesidad de información definida	-
Medición	Un conjunto de operaciones que tiene por objeto determiner el valor de una medida	-
Procedimiento de medición	Un conjunto de operaciones, descritos específicamente y usados en el desarrollo de una medición particular de acuerdo a un método dado.	-
Observación	Una instancia de aplicación de un procedimiento de medición para producir un valor para una medida base.	-
Tipo de método	El tipo de método depende de la naturaleza de la operación usada para cuantificar un atributo. Se pueden distinguir dos tipos de métodos: los subjetivos (con una cuantificación que implica el juicio humano) y los objetivos (con una cuantificación basada en reglas numéricas)	-
Valor	Un resultado numérico o categórico asignado a una medida base, una derivada o un indicador – un estadístico.	Métrica, Medida

Tabla 7. Conceptos de ISO/IEC 15939 usados para dar respuesta a las preguntas "cómo medir" y cuántos datos son necesarios. [20]

aspecto. Según la bibliografía de DQ, un método típico de medición de DQ para datos se puede calcular mediante la aplicación de una fórmula como la mostradas en la ecuación (1) [3, 22]:

(1) Ratio = 1 - [*NúmeroDeDatosQueNoSatisfacenUnCriterio/ NúmeroTotalDeDatos*]

La fórmula 1 representa una **función de medición** de una medida derivada. La medida está compuesta por dos medidas base: por un lado **NúmeroDeDatosQueNoSatisfacenUnCriterio** y por otro **NúmeroTotalDeDatos**. El método de medición para la primera es objetivo: consiste simplemente en contar el número de datos que no cumplen un determinado criterio. Este criterio suele ser venir

Elemento DQMIM	Concepto Medible	
	Compleción	Fiabilidad
Método de Medición	*"Calcular la tasa de noticias (elementos) que tienen valor para todos los ítems definidos"*	*"Calcular la tasa de noticias que tienen una fuente fiable"*
Domain of values	[0, 1]	[0, 1]

Tabla 8. Conceptos medibles para el ejemplo.

dado por una regla de negocio [11, 24, 41]. El resultado de decidir si la unidad de datos cumple el criterio puede ser "*Verdadero*" o "*Falso*". Así que, con el fin de obtener un valor para la medida de **NúmeroDeDatosQueNoSatisfacenUnCriterio**, se debe realizar un recuento del número de juicios de los datos que hayan obtenido un valor de "*verdadero*". La verdadera dificultad reside tanto en enumerar la regla de negocio como en decidir si un dato cumple o no la cumple. Para determinar si se cumple o no la regla de negocio es preciso determinar objetiva o subjetivamente si un valor para un dato pertenece realmente al dominio correspondiente de ese dato. En la mayoría de las ocasiones, se puede realizar el juicio sobre el mismo valor del dato; pero a veces, es necesario proporcionar junto con el valor, otro(s) valor(es) que complemente(n) el significado del valor que se va a juzgar en el sentido apuntado por el concepto medible. En la bibliografía, a estos nuevos valores se les suele llamar *metadatos*, aunque nosotros pensamos que este término no es del todo correcto porque si bien dichos *metadatos* aportan información sobre el dato, en otras oca-

	Juicio Objetivo	Juicio Subjetivo
Valores Objetivos	El estudio de la actualidad para una oferta de empleo publicada en Internet con unas fechas de inicio y de finalización de vigencia. El sitio web proporciona ambos datos al usuario, quien puede hacer un juicio objetivo de la vigencia de la oferta, mediante la comparación entre el rango dado y la fecha actual y decidir si está a tiempo de inscribirse a la oferta del trabajo (independientemente de lo buenas que sean las condiciones del trabajo o el propio trabajo)	El estudio de valor añadido de un dato: e.g. alguien que quiera comprar una cámara de fotos digital puede estar interesado sólo en ciertas características técnicas específicas, como si la cámara tiene o no zoom óptico. Si el fabricante ofrece datos sobre el tipo de alimentación (un valor objetivo), el valor añadido de los datos sobre la cámara es sin duda mayor. Ahora, el usuario puede decidir si el nuevo valor proporcionado hacer mejor o no la información sobre la cámara (no si la cámara es mejor o peor)

Tabla 9. Ejemplos de situaciones en la que la calidad de los datos se mide haciendo una comparación con un valor dado que se usa como referencia

siones no es relevante para describir el hecho en sí. Por esta razón pensamos que el nombre adecuado debería ser el de *extradatos* (englobaríamos en este conjunto aquellos que en la bibliografía han sido llamados *metadatos* con este mismo significado). No obstante, y para no separarnos de la bibliografía, hemos decidido seguir usando el término *metadato*. [26] identifica como posibles fuentes para valores de estos metadatos los siguientes elementos: un implicado en el proceso de medición, el proceso de fabricación de datos o incluso el mismo almacén de datos. Diferentes autores en el campo de DQ están de acuerdo en que los valores de los metadatos procedentes de un usuario son probablemente subjetivos, mientras que los procedentes de los almacenes de datos pueden considerarse en su generalidad objetivos. La Tabla 9 muestra varios ejemplos de situaciones con diferentes tipos de juicios.

Creemos que es muy interesante la identificación hecha en [26] de los diferentes métodos para generar valores de metadatos de acuerdo con el concepto de medición para medir y la fuente de estos valores. Estos métodos han sido reproducidos en la Tabla 10.

La tenencia de metadatos facilita el proceso de evaluación de aquellos conceptos medibles en los que el propio valor del dato es insuficiente. En ocasiones, estos metadatos pueden formar ya parte del modelo de datos del almacén de datos (por ejemplo, cuando están almacenados como un atributo relacional más y que puede ser usado para este fin), pero en otras, el modelo de datos puede no recogerlo, por ejemplo, por ser demasiado explícito y tiene por tanto que ampliarse con el fin de almacenar dichos metadatos como se recomienda en [40] para el modelo relacional, o incluso como se propone en [6] para la Web Semántica. En la sección 3 se propone un esquema XML que permite almacenar los metadatos que no existen en el modelo de datos junto con datos que amplían.

Fuente del metadato	Concepto Medible	Método para generar los extradatos
Subjetivo (usuario)	Credibilidad	Experiencia del Usuario
	Interpretabilidad	Muestro del Usuario
	Reputación	Experiencia del Usuario
	Valor Añadido	Valoración por parte del usuario
	Soporte al Cliente	Análisis Sintático, Contrato
	Objectividad	Entrada de Usuario Experto
	Fiabilidad	Valoración por parte del usuario
	Oportunidad	Análisis Sintáctico

Tabla 10. Clasificación de los métodos para la generación de los metadatos de medición para evaluar Concepto [26]

CodNoticia	TextoDeLaNoticia	FechaProducciónNoticia	CodProveedorNoticia
N001	*"Se ha publicado un estándar que contiene un modelo estándar de calidad de datos"*	29/01//2010	MIT News.
N003	"El Primer Congreso Europeo de Calidad de Datos se celebrará en España en 2011"	25/01/2010	MyPersonalDQBlog.com

Tabla 11. Datos usados en el ejemplo (las noticias no son necesariamente verdad).

A veces, para realizar adecuadamente una evaluación se requiere algo más que un simple valor procedente de una medida de base o calculado mediante el uso de una función como parte de una medida derivada. Es necesario dar una interpretación cualitativa de los resultados de estas medidas. Esta interpretación podría hacerse a través de un **modelo de análisis** que permitiera determinar un valor representativo para un concepto de medición dado, en combinación con unos **criterios de decisión** para determinar si un dato es lo suficientemente bueno para ser usado en una tarea. Esta tipo de medida es conocida como **indicador**. Este término no puede ser confundida con el de "calidad de Indicadores" introducido en [39].

La relación que se puede establecer entre los términos que dan respuesta a esta pregunta es la siguiente: una "*medida*" puede ser de uno de estos tipos: "*medida base*", "*medidas derivadas*" (incluye una función que opera con otras medidas base o con otras medidas derivadas), o un "*indicador*" (incluye un modelo de análisis y un criterio de decisión). El modelo de análisis puede incluir otros tipos de medidas (bien indicadores, medidas base o medidas derivadas)

Análogamente a lo que se viene haciendo en secciones anteriores, se van a ilustrar los conceptos introducidos a través del ejemplo. Se va a representar el documento *RSS* con las noticias como si fuesen tuplas relacionales (véase Tabla 11).

CodProveedorNoticias	GradoFiabilidad
MIT News.	'Alto'
MyPersonalDQBlog.com	'Bajo'

Tabla 12. Grado de confiabilidad para cada proveedor de noticias

Para medir la compleción, el método de medición correspondiente que se va a definir consiste en calcular la proporción de elementos (tuplas según la representación de la Tabla 11) que no tienen valores nulos. Esta es una medida derivada que puede calcularse aplicando la siguiente función (2):

(2) Compleción (NoticiasEnDocumentoRSS) = 1 - NúmeroNoticiasNoCompletas/NúmeroDeNoticias

La función se compone de dos medidas de base *NúmeroNoticiasNoCompletas* y *NúmeroDeNoticias*. La segunda, *NúmeroDeNoticias,* es una medida base cuyo valor puede ser obtenido contando el número de noticias. Es obvio que para el ejemplo, el resultado es 3. El primero, *NúmeroNoticiasNoCompletas*, puede calcularse como el conteo de los resultados de un juicio objetivo que toma la siguiente regla de negocio: "*si una noticia tiene valores nulos entre sus atributos, entonces no es completa, de lo contrario es completa*". La aplicación de esta regla lleva a ver que como la segunda tupla (correspondiente a la noticia con *CodNoticia* N002 tiene el atributo *FechaProducciónNoticia* a *null,* entonces se puede determinar que el resultado es 1. La aplicación de la función de medición (2) da el valor de 1-1/3= 0.667 (66,7% de las noticias son completas)

Para evaluar la fiabilidad de las noticias, se puede usar la fórmula (3):

(3) Fiabilidad (NoticiasEnDocumentoRSS) = 1-NúmeroNoticiasNoFiables / NúmeroDeNoticias

Al igual que antes, esta medida base tiene dos medidas derivadas. Ahora, la regla que muestra cómo determinar si una noticia es o no fiable se puede enunciar como sigue: "*si una noticia ha sido proporcionada por un proveedor de noticias con un grado de fiabilidad 'bajo', entonces no es fiable, de lo contrario es fiable*". En este caso es necesario un metadato (grado de fiabilidad asociado a un proveedor de noticias) que no aparece en el modelo de datos y que es necesario proporcionar. Este grado de fiabilidad viene representado por el atributo *GradoDeFiabilidad* y se le han asignado los valores mostrados en Tabla 12. Estos valores, en este ejemplo pueden ser considerados totalmente subjetivos.

Al efectuar los juicios correspondiente sobre las noticias de la Tabla 11, usando los metadatos proporcionados en la Tabla 12 y aplicando la fórmula (3), es fácil comprobar que se obtiene el valor de 0,334 para la fiabilidad (es decir, sólo el 33,4% de las noticias son fiables).

A fin de ilustrar a través de nuestro ejemplo el concepto de **indicador**, se va a introducir el siguiente, ***NivelDQDeNoticiasPublicadas***, que es una medida cuyo objetivo es satisfacer la necesidades de información para nuestro ejemplo enunciada en la sección 2.2. Así puede decirse que el objeto de este indicador es determinar el nivel de calidad de los datos en las noticias proporcionadas por el periódico a sus clientes (es importante tener en cuenta que esta percepción de la calidad puede no ser compartida por los clientes, sino que está siendo evaluada desde el punto de vista de los editores). El modelo de análisis depende de los dos conceptos medibles y puede ser enunciado: ***NivelDQDeNoticiasPublicadas (integridad, fiabilidad)***. El criterio de decisión se muestra en la Tabla 13. Los rangos de valores de aceptación mostrados en dicha tabla son ejemplos propuestos como umbrales proporcionados por los editores, y deben ser tomados como ejemplo.

Como hemos obtenido un valor de 0,667 para la compleción y 0,334 para la fiabilidad, se puede concluir a partir del criterio de decisión que las noticias tienen un nivel de calidad de datos "**Bajo**".

Al medir los valores de datos, a veces puede resultar casi imposible o preferible no evaluar todos los datos existentes, debido a varias razones, como por ejemplo para evitar un excesivo coste computacional que sature el rendimiento del sistema de información. En estas situaciones, puede ser interesante realizar un muestreo de los datos. La muestra obtenida debe ser representativa de la población de los datos, de modo que pueda ser posible tener éxito cuando se extrapolen los resultados, como se señala en [34] y en [24].Varios autores se han hecho eco de esta necesidad y han trabajado en esta área: por ejemplo en [11] se propone extraer muestras aleatorias de datos, pero con la misma probabilidad de ser elegidos; los mismos autores también proponen varias directrices de diseño de muestreo de los datos para las diferentes necesidades de información. Por su parte, en [22] se explica cómo determinar el tamaño de la muestra con el fin de acotar la cantidad de errores. En [5] se sugiere usar el estándar ISO 2859 [18] para calcular los parámetros del muestreo. Aunque el estándar ISO/IEC 15939 no incluye propiamente el muestreo, teniendo en cuenta los antecedentes mencionados, hemos decidido incluirlo en este DQMIM. Así que se introducen dos nuevos conceptos "***tamaño de la muestra***" y "***método de***

		Compleción	
		[0, 0.8)	[0.8, 1]
	[0.6, 1]	"Aceptable"	"Alto"

Tabla 13. Criterios de decisión para NivelDQDeNoticiasPublicadas (los valores se proponen como ejemplo)

muestreo". En nuestro ejemplo de trabajo, el número de noticias es suficientemente pequeño como para hacer los cálculos con un reducido coste computacional (de hecho han sido realizados a mano), así que no es necesario tomar muestras de la población de datos.

En la Tabla 12 se mostraron unos valores que realmente definían una regla de negocio, que representa la percepción de la organización del grado de fiabilidad de cada proveedor de noticias. Estos valores, aunque en este caso se han propuesto como ejemplo, podrían ser proporcionados sólo por una persona con suficiente autoridad sobre el negocio, o bien haber sido agregados a partir de las opiniones al respecto procedentes de varias personas. En cualquier caso, los criterios utilizados por una persona para, por ejemplo, decidir que el proveedor de datos "*MIT News*" tiene un alto grado de fiabilidad, se basa en sus connotaciones personales hacia esta fuente de datos; estas connotaciones han sido determinadas por su propia experiencia, por influencias externas, por su particular percepción de la calidad,...

Como no hay una manera objetiva de manejar estas connotaciones, pues están cargadas de una gran subjetividad

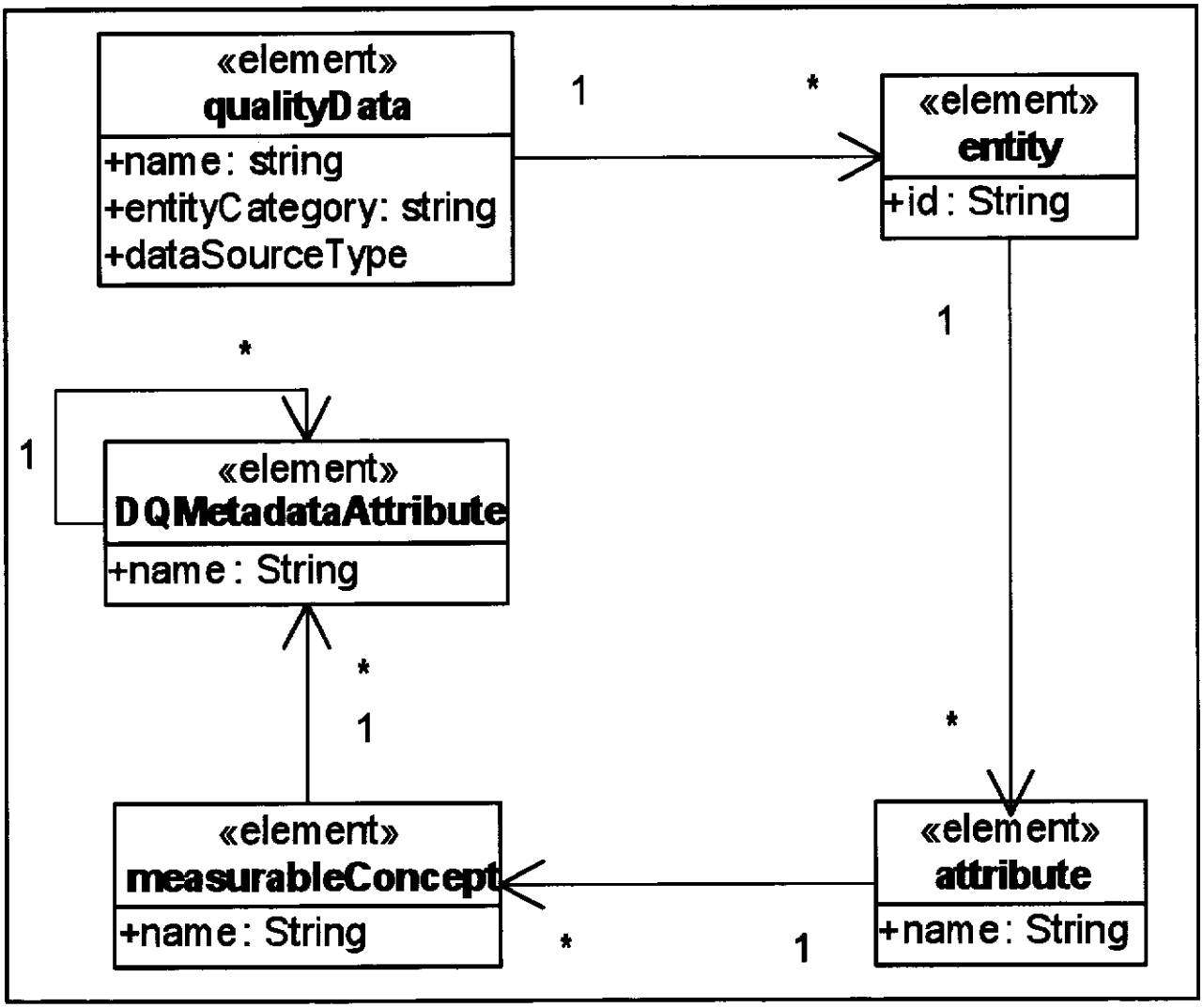


Figura 1. DQ XSD for storing entity metadata.

y por tanto de incertidumbre, algunos autores consideran necesario introducir y manejar esas connotaciones usando los fundamentos y métodos proporcionados por la lógica difusa, como los operadores OWA [7, 17]. Por ejemplo, supóngase que se ha celebrado una reunión de todos los editores (gestores de calidad de datos) que participan en nuestro ejemplo, y que en dicha reunión van a realizar una votación sobre el valor del grado de fiabilidad de cada proveedor de noticias; tras haber aportado cada uno su opinión en la forma {"alto", "medio", "bajo"}, se necesita agregar todas las opiniones con el fin de obtener un valor representativo que sintetice todas estas opiniones. Obviamente, el precio que hay que pagar por gestionar esa subjetividad, es un incremento de la complejidad en la definición del método de medida.

3.5 ¿CUÁNDO MEDIR?

Aunque ISO/IEC 15939 no identifica cuando se debe realizar una medición, es una cuestión realmente importante que merece la pena ser incluida en este estudio. En diferentes momentos de su ciclo de vida [35], los datos son utilizados por trabajadores que pueden estar desempeñando diferentes roles, y que por tanto pueden tener requisitos distintos de calidad de datos para cada una de las tareas que realizarán en momentos diferentes. En [11], se sugiere identificar adecuadamente el valor de la información en cada instante de la correspondiente cadena de valor.

Por otro lado, en [24] se hace una mención específica al aspecto temporal de la medición, distinguiendo entre un modo estático y otro dinámico. Hemos interpretado el modo dinámico como un conjunto de medidas estáticas realizadas en diferentes puntos del ciclo de vida de los datos. Podría ser como tomar instantáneas diferentes de los datos para realizar el seguimiento de los diferentes niveles de calidad de datos a través del sistema de información para una determinada tarea. En [34], se recomienda identificar el momento en que la medición se debe realizar con el fin de minimizar costes en términos

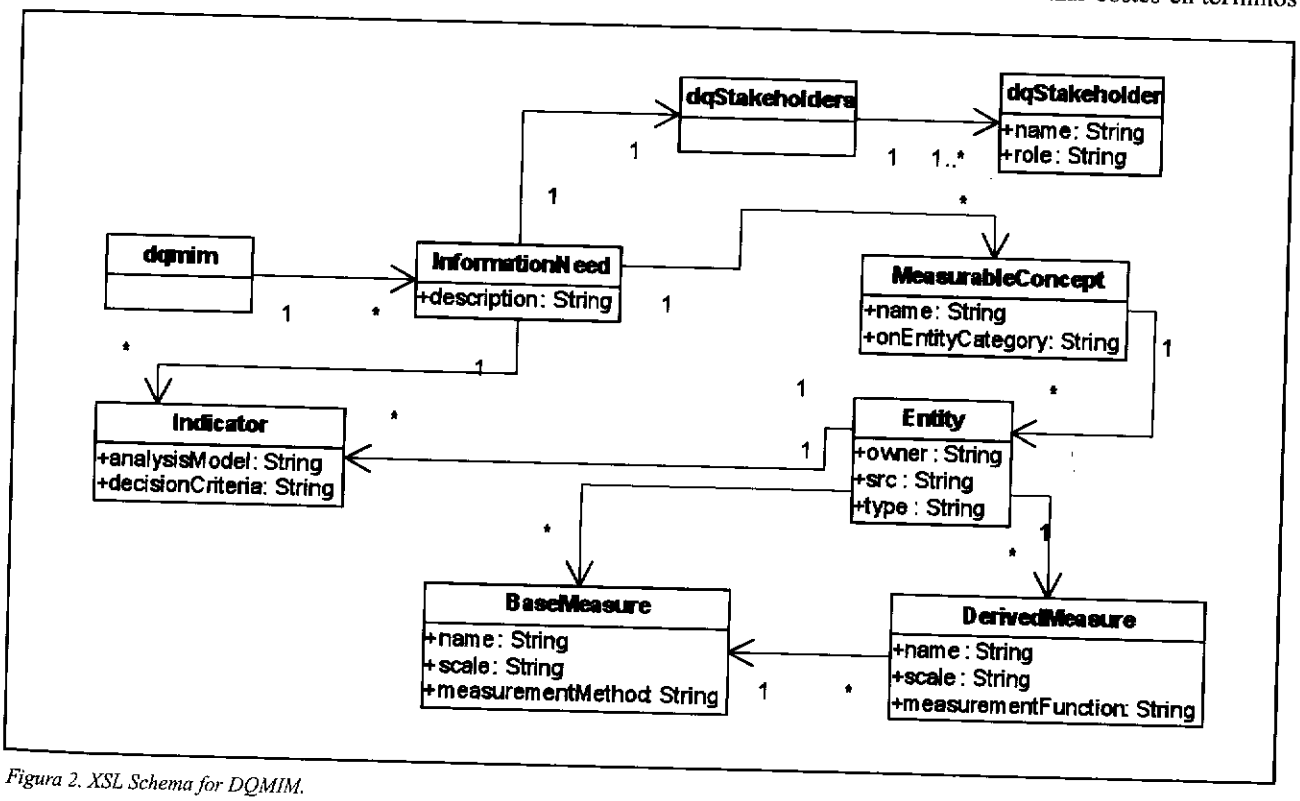


Figura 2. XSL Schema for DQMIM.

```
<qualityData name="News" entityCategory= "dataValues" dataSourceType= "text/XML">
        <entity id="N001">
            <attribute name="TextoDeLaNoticia"> "Se ha publicado un estándar que contiene un modelo estándar de calidad de
            datos" </attribute>
            <attribute name="FechaProducciónNoticia">29/01/2010</attribute>
            <attribute name= "CodProveedorNoticia">MIT News
                <measurableConcept name="Fiabilidad">
                    <DQMetadataAttribute name= "GradoFiabilidad">"Alto"</DQMetadataAttribute>
                </measurableConcept>
            </attribute>
        </entity>
        <entity id="N002">
            <attribute name=" TextoDeLaNoticia">"  El SEI introduce las mejores prácticas de gestión de calidad de datos en la
            Versión 2.0 de CMMI" </attribute>
            <attribute name= "FechaProducciónNoticia"></attribute>
            <attribute name= "CodProveedorNoticia">The Data News
                <measurableConcept name="Fiabilidad">
                    <DQMetadataAttribute name= "GradoFiabilidad">"Bajo"</DQMetadataAttribute>
                </measurableConcept>
            </attribute>
        </entity>
        <entity id="N003">
            <attribute name=" TextoDeLaNoticia"> "El Primer Congreso Europeo de Calidad de Datos se celebrará en España en
            2011" </attribute>
            <attribute name= "FechaProducciónNoticia">25/01/2010</attribute>
            <attribute name= "CodProveedorNoticia">MyPersonalDQBlog.com
                <measurableConcept name="Fiabilidad">
                    <DQMetadataAttribute name= "GradoFiabilidad">"Bajo"</DQMetadataAttribute>
                </measurableConcept>
            </attribute>
        </entity>
</qualityData>
```

Tabla 14. DQTaggedNews.XML: A DQXML containing data and metadata for the working example.

de recursos. Así que, para tener una idea completa de lo que representa cada medición, es necesario contextualizar la medición en un momento dado del ciclo de vida de los datos. Por ejemplo, en el ejemplo que se viene mostrando en este artículo, los conceptos medibles se miden antes de que los datos se distribuyan a los consumidores.

4. UN ESQUEMA XML PARA EL ALMACENAMIENTO DE METADATOS.

Como se dijo en la sección 2.5, cada uno de los implicados que necesitan desarrollar o usar medidas de calidad de datos sobre los valores de datos, precisan tener a su disposición y de forma accesible el conjunto de datos cuya calidad desea medirse. Es probable que estos datos deban ser completados con algunos metadatos que complementen el significado del dato en la dirección del concepto medible; esto implica identificar los metadatos en el modelo de datos si existen, o añadiéndolos si no existieran. Teniendo en cuenta que los diferentes roles pueden requerir diferentes valores de metadatos para cada proceso de medición (por ejemplo, si para medir la misma dimensión de calidad se necesitan metadatos diferentes), parece razonable permitir que un implicado tenga disponibles para el plan de medición específico, los metadatos específicos definidos para su método de medición enlazados de alguna forma a los datos que quiere evaluar. Además, puede ser necesario como se recomienda en [28], que las medidas realizadas sobre la calidad de los datos se guarden también en el modelo de datos (junto a los correspondientes metadatos), a fin de facilitar los correspondientes análisis posteriores o simplemente reutilizarlas para comparar la eficiencia de acciones correctoras.

```
<dqmim>
<InformationNeed>
        <description>
            Saber si el nivel de DQ de las noticias publicadas en una Web satisface lo suficiente a los visitants como para que
            vuelvan.
        </description>
        <dqStakeholders>
            <dqStakeholder name= "Ismael Caballero" role= "Líder del Proceso de Medición"</dqStakeholder>
            <dqStakeholder name= "Mario Piattini" role="Proveedor de Datos"</dqStakeholder>
        </dqStakeholders>
        <MeasurableConcept id= "Compleción" onEntityCategory= "DataValues">
            <entity type= "text/XML" src= "DQTaggedNews.XML">
                <owner>Coral Calero</owner>
                <BaseMeasure name= "NúmeroDeNoticiasNoCompletas" scale= "Ratio">
                    <MeasurementMethod>
                        <description>Calcular la tasa de noticias (elementos) que tienen al menos un valor para todos los
                        elementos definidos </description>
                    </MeasurementMethod>
                </BaseMeasure>
                <BaseMeasure name= "NúmeroDeNoticias" scale= "Ratio">
                    <MeasurementMethod>
                        <description>Número de Noticias en un document RSS </description>
                    </MeasurementMethod>
                </BaseMeasure>
                <DerivedMeasure name= "Compleción">
                    <BaseMeasure id= "NúmeroDeNoticiasNoCompletas"/>
                    <BaseMeasure id= "NúmeroDeNoticias"/>
                    <MeasurementFunction> 1- NúmeroDeNoticiasNoCompletas / NúmeroDeNoticias
                    </MeasurementFunction>
                </DerivedMeasure>
            </entity>
        </MeasurableConcept>
        <MeasurableConcept id= "Fiabilidad" onEntityCategory= "DataValues">
            <entity type= "text/XML" src="DQTaggedNews.XML">
                <owner>Coral Calero</owner>
                <BaseMeasure name= "NúmeroDeNoticiasNoFiables" scale= "Ratio">
                    <MeasurementMethod>
                        <description> Calcular la tasa de noticias que tienen una fuente fiable </description>
                    </MeasurementMethod>
                </BaseMeasure>
                <BaseMeasure name= "NúmeroDeNoticias" scale= "Ratio">
                    <MeasurementMethod>
                        <description>Número de noticias en el document RSS </description>
                    </MeasurementMethod>
                </BaseMeasure>
                <DerivedMeasure name= "Fiabilidad">
                    <BaseMeasure id= "NúmeroDeNoticiasNoFiables"/>
                    <BaseMeasure id= "NúmeroDeNoticias"/>
                    <MeasurementFunction> 1-NúmeroDeNoticiasNoFiables / NúmeroDeNoticias </MeasurementFunction>
                </DerivedMeasure>
            </entity>
        </MeasurableConcept>
```

Tabla 15. Fichero XML basado en DQMIM-XSD correspondiente a nuestro ejemplo.

A fin de conseguir mantener enlazados valores con metadatos, en [40], los modelos relacionales de los correspondientes modelos de datos se extienden con atributos dedicados a almacenar el valor para los metadatos necesarios para la medición de un concepto medible. Este proceso se conoce como "*etiquetar*". Dado que un atributo relacional es la unidad mínima básica del modelo relacional, se recomienda etiquetar los datos a este nivel de granularidad. Estas operaciones se basan en la noción de lo que Wand et al. llamaron en [38] "*indicador de calidad*", y que se corresponde con lo que nosotros venimos llamando "*metadatos*". En la propuesta realizada en [40], se desarrollan mecanismos para facilitar el enlace entre un atributo del almacén de datos y su(s) correspondiente(s) metadato(s), a través del concepto "*clave de calidad*". En el modelo relacional, este enlace se puede realizar mediante el uso de subrogados.

Puesto que el XML es actualmente la tecnología de preferencia para el intercambio de datos, se ha pensado en utilizarla para enlazar datos con sus correspondientes metadatos para cada rol. Para ello, se ha diseñado un esquema XML al que se ha llamado DQXSD. Los datos pueden provenir de cualquier almacén de datos. Los metadatos pueden haber sido generados usando cualquiera de los métodos propuestos en la Tabla 10. El documento obtenido recibirá el nombre de documento DQXML.

Como puede verse en la Figura 1, el elemento principal es ***qualityData***, la raíz que agrupa a los valores de datos. Se compone de una secuencia de elementos ***entity***. Cada ***entity*** modela una entidad (por ejemplo, una tupla de una base de datos relacional, o un elemento de un documento XML). Una ***entity*** puede estar compuesta por un conjunto de elementos ***attribute,*** que contienen los valores del atributo. Hasta aquí, lo único que se ha hecho es asignar los valores de datos desde su origen a una nueva estructura de datos. La novedad viene aquí: el ***attribute*** se amplía ahora con un nuevo elemento ***measurableConcept***, que se destina a guardar más información sobre concepto medible para el que se requieren los metadatos. Como cada concepto medible puede necesitar diferentes metadatos, se propone un nuevo elemento ***DQMetadataAttribute*** que se

```
        <Indicator>
            <analysisModel>
                <description> NivelDQDeNoticiasPublicadas (Compleción, Fiabilidad) </description>
            </analysisModel>
            <decisionCriteria>
                <value label= "ALTO">
                    <Measure= "Compleción" fromClosed= "0.8" toClosed= "1">
                    <Measure= "Fiabilidad" fromClosed= "0.6" toClosed= "1">
                </value>
                <value label= "ACEPTABLE">
                    <Measure= "Compleción" fromClosed= "0.8" toClosed= "1">
                    <Measure= "Fiabilidad" fromClosed= "0" toOpen= "0.6">
                </value>
                <value label= "ACEPTABLE">
                    <Measure= "Compleción" fromClosed= "0" toOpen= "0.8">
                    <Measure= "Fiabilidad" fromClosed= "0.6" toClosed= "1">
                </value>
                <value label= "BAJA">
                    <Measure= "Compleción" fromClosed= "0" toOpen= "0.8">
                    <Measure= "Fiabilidad" fromClosed= "0" toOpen= "0.6">
                </value>
            </decisionCriteria>
        </Indicator>
</InformationNeed>
</dqmim>
```

Tabla 16. Fichero XML basado en DQMIM-XSD correspondiente a nuestro ejemplo. [Continuación].

introduce en el esquema XML para el almacenamiento de los metadatos correspondientes.

Como ejemplo de aplicación, se han representado los datos mostrados en formato relacional en la Tabla 11, así como los metadatos necesarios mostrados en la Tabla 12 correspondientes al ejemplo que se viene mostrando a lo largo de este trabajo. El resultado es el fichero DQTaggedNews.XML mostrado en la Tabla 14.

5. DQMIM-XSD: A XML SCHEMA FOR DQMIM

El objetivo principal de esta sección es reunir de forma resumida todos los conceptos presentados en la sección 2. Con este fin, se ha definido un esquema XML que permite representar estos conceptos y sus relaciones en un documento XML. Este esquema ha sido llamado DQMIM-XSD (véase la estructura en la Figura 2).

El elemento raíz es ***dqmim*** que integra todos los conceptos. Se compone de un conjunto de elementos del tipo ***InformationNeed*** (sección 2.2) que representa las diferentes necesidades información que motivan el proceso de medición. Una ***InformationNeed*** contiene un grupo de ***dqStakeholders*** (sección 2.4), elemento que sirve para especificar el nombre de los implicados y su rol en el proceso de medición. Una colección de elementos del tipo ***MeasurableConcept*** (sección 2.3) se utilizan para especificar los diferentes conceptos medibles necesarios para satisfacer una ***InformationNeed***. Cada ***MeasurableConcept*** se mide sobre una ***entity*** (sección 2.3). La medida resultante puede ser una ***BaseMeasure*** o una ***DerivedMeasure*** (sección 2.5). En una ***BaseMeasure***, se utiliza un método de medición, mientras que en una ***DerivedMeasure***, se usa una función de medición. Por último, se especifican los elementos del tipo ***indicator*** necesarios (sección 2.5).

La Tabla 15 y Tabla 16 muestran un documento DQMIM-XML generado como una instanciación de DQMIM-XSD para nuestro ejemplo. Es preciso tener en cuenta que los datos y metadatos que se usan en la medición son los almacenados en el documento DQXML llamado DQTaggedNews.XML (véase Tabla 14).

6. CONCLUSIONES Y TRABAJO FUTURO

Tener calidad en los datos que maneja una organización es un hecho diferencial hoy día. La facilidad y el bajo coste de adquirir nuevos datos permiten la posibilidad de tener más datos de los necesarios, lo que en determinada circunstancias lleva a situaciones indeseables en las que el exceso de datos redundantes e inútiles perjudica a las organizaciones.

Por esta razón las organizaciones deben implementar mecanismos para la gestión de la calidad de los datos contenidos en sus recursos organizacionales. Teniendo en cuenta la máxima "*no se puede gestionar lo que no se puede medir*", el primer paso en cualquier iniciativa de gestión de calidad de datos debe ser la definición de medidas de calidad de datos sobre esos recursos de datos. Para lograr este objetivo es necesario tener en cuenta la naturaleza especial de los datos. Esta naturaleza condiciona numerosos aspectos de la medición. Estos aspectos de la medición han sido tratados por diferentes autores del área, pero de una forma tan específica a su contexto, que a veces no es posible ni reutilizar la terminología que han definido. A lo largo de este artículo se han analizado estos aspectos de medición y se han unificado en torno a un Modelo de Información de Medición de Calidad de Datos (DQMIM). Este DQMIM propone una terminología común para los conceptos que deben usarse cuando se diseñe un Plan para la Medición de Calidad de Datos.

El objetivo de nuestra investigación está puesto en la estandarización al mayor nivel posible de los elementos relacionados con la medición. Una vez conseguida la unificación de términos, nuestro siguiente objetivo es intentar unificar y parametrizar diferentes métodos de medición, así como desarrollar mecanismos que permitan expresar dichos métodos de medición de tal manera que puedan ser utilizado entre distintas aplicaciones para distintos

datos, lo que hemos dado en llamar, interoperabilidad semántica de mediciones de calidad de datos.

7. AGRADECIMIENTOS

Esta investigación es parte de los proyectos PEGASO (TIN2009-13718-C02-01), ENGLOBAS (PII2109-0147-8235) y de la Red DQNet (TIN2008-04951-E/TIN)

8. REFERENCIAS

[1] D. P. Ballou, *et al.*, "Modelling Information Manufacturing Systems to Determine Information Product Quality," Management Science, vol. 44,April 1998 1998, pp. 462-484.

[2] C. Batini, *et al.*, *Conceptual Database Design - An Entity-Relationship Approach*: Benjamin/Cummings, 1992.

[3] C. Batini and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*. Berlin: Springer-Verlag Berlin Heidelberg, 2006.

[4] M. S. E. Burgess, *et al.*, "Quality Measures and the Information Consumer," in *Ninth International Conference on Information Quality (ICIQ'04)*, MIT, Cambridge, MA, USA, 2004, pp. 373-388.

[5] I. Caballero, *et al.*, "MMPro: A Methodology Based on ISO/IEC 15939 to draw up Data Quality Measurement Processes.," in *XIII International Conference on Information Quality*, MIT, Cambridge, MA, USA, 2008, pp. 326-340.

[6] I. Caballero, *et al.*, "DQRDFS:Towards a Semantic Web Enhanced with Data Quality," in *Web Information Systems and Technologies*, Funchal, Madeira, Portugal, 2008, pp. 178-183.

[7] I. Caballero, *et al.*, "Tailoring Data Quality Models using Social Network Preferences," in *2nd International Workshop on Managing Data Quality in Collaborative Information Systems* Brisbane, Australia, 2009, pp. 1-15.

[8] C. Calero, *et al.*, "Empirical Validation of Referential Integrity Metrics," Information and Software Technology. Special Issue on Controlled Experiments in Software Engineering, vol. 43,2001, pp. 949-957.

[9] C. Cappiello, *et al.*, "Data quality assessment from the user's perspective," in *International Workshop on Information Quality in Information Systems, (IQIS2004)*, Paris, Francia, 2004, pp. 68-73.

[10] A. Caro, *et al.*, "A proposal for a set of attributes relevant for Web Portal Data Quality," Software Quality Journal, March 15, 2008 2008,

[11] L. English, *Improving Data Warehouse and Business Information Quality: Methods for reducing costs and increasing Profits*. New York, NY, USA: Willey & Sons, 1999.

[12] M. Eppler, *Managing Information Quality: Increasing the Value of Information in Knowledge-intensive Products and Processes*: Springer, 2003.

[13] A. Even and G. Shankaranarayanan, "Value-Driven Data Quality Assessment," in *Tenth International Conference on Information Quality (ICIQ'05)*, MIT, Cambridge, MA, USA, 2005, pp. 13-26.

[14] M. Gebauer, *et al.*, "Reproducible Measurement of Data Quality Field," in *Tenth International Conference on Information Quality (ICIQ'05)*, MIT, Cambridge, MA, USA, 2005, pp. 237-246.

[15] M. Genero, *et al.*, "A survey of Metrics for UML Class Diagrams," Journal of Object Technology, vol. 4,2005, pp. 59-92.

[16] M. Gertz, *et al.*, "Report on the Dagstuhl Seminar "Data Quality on the Web"," SIGMOD RECORD, vol. 33,2004, pp. 127-132.

[17] E. Herrera-Viedma, *et al.*, "Evaluating the Information Quality of Web Sites: A Quality Methodology Based on Fuzzy Computing with Words," Journal of American Society for Information Science and Technology, vol. 54,2006, pp. 538-549.

[18] ISO, "ISO 2859-1: Sampling procedures for inspection by attributes -- Part 1: Sampling schemes indexed by

[35] T. C. Redman, *Data Quality for the Information Age*. Boston, MA, USA: Artech House Publishers, 1996.

[36] D. M. Strong, *et al.*, "Data Quality in Context," Comm. of the ACM, vol. 40,May 1997 1997, pp. 103-110.

[37] W3C. (2007, *Extensible Markup Language (XML).*

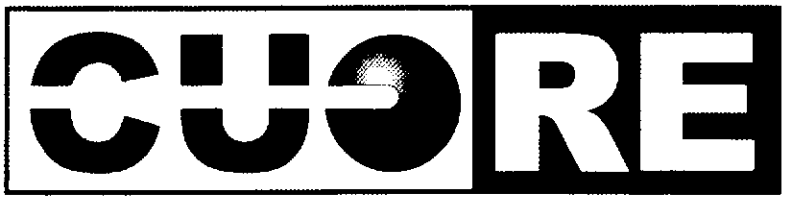
[38] Y. Wand and R. Y. Wang, "Anchoring Data Quality Dimensions in Ontological Foundations," Communications of the ACM, vol. 39,November 1996 1996, pp. 86-95.

[39] R. Y. Wang and S. Madnick, "Data Quality Requirements: Analysis and Modelling," in *Ninth International Conference on Data Engineering (ICDE'93)*, Vienna, Austria, 1993, pp. 670-677.

[40] R. Y. Wang, *et al.*, "Towards quality data: An attribute-based approach," Journal of Decision Support Systems, vol. 13,1995, pp. 349-372.

[41] R. Y. Wang, "A Product Perspective on Total Data Quality Management," Comm. of the ACM, vol. 41,February 1998 1998, pp. 58-65.

[42] R. Y. Wang, "A Product Perspective on Total Data Quality Management," Communications of the ACM, vol. 41,February 1998 1998, pp. 58-65.


CUORE