Una Metodología Basada en ISO/IEC 15939 para la Elaboración de Planes de Medición de Calidad de Datos

Eugenio Verbo[1], Ismael Caballero[12], Ricardo Pérez[1], Coral Calero[2], Mario Piattini[2]

[1] Indra Software Labs, Departamento de I+D de Indra Software Labs,
Ronda de Toledo s/n 13004 Ciudad Real, España
{emverbo, icaballerom, rpdelcastillo}@indra.es
[2] Grupo de Investigación ALARCOS, Instituto de Tecnologías de la Información y Sistemas,
Paseo de la Universidad 4, 13071, España
{Ismael.Caballero, Coral.Calero, Mario.Piattini}@uclm.es

Resumen. Hoy en día, los datos juegan un papel fundamental en las organizaciones y la gestión de su calidad se está convirtiendo en una actividad imprescindible. Como parte de dicha gestión, y en aras de obtener medidas útiles, las organizaciones necesitan realizar planes de medición de Calidad de los Datos (CD). Estos planes de medición deben hacerse teniendo en cuenta la propia naturaleza de los datos así como los factores organizacionales que afectan al uso de los mismos. Dado que no existen muchos trabajos en la bibliografía con este objetivo, este artículo presenta una metodología, MEPLAMECAL, para elaborar planes de medición de CD. MEPLAMECAL está basada en ISO/IEC 15939 que, a pesar de ser un estándar de calidad de software, consideramos que puede ser aplicado en este contexto debido a las similitudes existentes entre software y datos. La metodología propuesta está compuesta por dos actividades: (1) establecer y mantener el compromiso de la organización en el proceso de medición, y (2) la elaboración del plan de medición. Estas dos actividades se estructuran en tareas para las que se proponen productos de entrada y salida, así como técnicas y herramientas a utilizar, muchas de ellas tomadas de la Ingeniería del Software.

Palabras Clave: Calidad de Datos, Medición de la Calidad de Datos, ISO/IEC 15939, Modelo de Información de Medición de la Calidad de Datos.

1 Introducción

Levis et al. en [26] mencionan ejemplos de escenarios en los que datos con niveles de calidad inadecuados originan problemas que afectan negativamente a los Sistemas de Información y, por tanto, al rendimiento organizacional. Las causas más comunes de esos niveles inadecuados son una serie de obstáculos a lo largo del ciclo de vida de los datos en el SI como los descritos por Strong et al. en [33]. Por su parte, los problemas producidos por esta falta de calidad pueden ser clasificados a diferentes niveles según su naturaleza: técnicos (como los referidos a la implementación de almacenes de datos [29]), organizacional (tales como pérdida de clientes [31], grandes pérdidas financieras [13, 27] o incluso insatisfacción de los trabajadores [12, 33]) y legales (como la violación del artículo cuarto del Apartado II de la LOPD de 1999).

Con el objetivo de minimizar el impacto negativo de estos problemas en el desarrollo de sus actividades, es fundamental que las organizaciones sean capaces de evaluar si el nivel de CD de sus datos es el adecuado. Esto implica por un lado la definición de medidas de CD sobre ellos y por otro el establecimiento de unos rangos de aceptación válidos para los valores obtenidos en esas medidas. De esta forma, y aplicando las técnicas y herramientas clásicas de calidad será posible localizar de forma más eficiente los datos con niveles de CD inadecuados y sus causas.

Sin embargo, debido a la propia naturaleza de los datos, puede ser bastante difícil realizar una definición adecuada de las medidas. Una aproximación que puede hacer más sencillo el proceso de definición de medidas, es la propuesta por el Programa TDQM del MIT en [32, 37], en la que se aporta una visión desde las teorías clásicas de calidad, en la que los datos pueden ser considerados como la materia prima de un proceso de producción en el que la información es el producto resultante. Teniendo en cuenta esta analogía de información con los clásicos productos manufacturados se habilita la aplicación de los principios clásicos de gestión de la calidad a la CD como algunos autores, de forma implícita o explícita, habían sugerido ([3, 6, 26]).

La bibliografía consultada describe algunas metodologías genéricas de evaluación de CD como la propuesta por Lee et al. en [25] (donde no se describen los procedimientos de medición como parte de la definición de las medidas), y otras particulares centradas en los problemas específicos de los escenarios donde se requería la medición como por ejemplo la propuesta por Al-Hakim en [1] para entornos médicos, donde habitualmente las medidas son realizadas *ad hoc*, sin una planificación previa y casi siempre sin tener en cuenta algunos aspectos importantes relativos tanto a la disponibilidad y disposición de los recursos organizacionales para el proceso de medición como a la naturaleza específica de los datos que afectan al propio proceso de medición. Desafortunadamente, esta suele ser una carencia común en los trabajos encontrados.

No obstante, algunos trabajos como los propuestos en [4, 17, 19, 28] analizan varios de los aspectos mencionados, y tratan de generalizar y priorizar sus conclusiones. Sin embargo, y a pesar de haber obtenido un conjunto de conceptos con bastantes elementos comunes, se da la circunstancia de que normalmente han usado diferentes términos para el mismo concepto, lo que dificulta la puesta en práctica de los resultados que se han obtenido en escenarios distintos de los suyos. Para tratar de paliar esta inconsistencia en la nomenclatura, Caballero et al. en [7] han analizado los términos de los autores más referenciados y han propuesto una terminología unificada de acuerdo con el Modelo de Información de Medición de Software propuesto en ISO/IEC 15939 [23] y la han ampliado con los aspectos específicos de CD . El resultado es un **Modelo de Información de Medición de Calidad de Datos** (*Data Quality Measurement Information Model*, **DQMIM**).

Un valor añadido de ISO/IEC 15939 es que proporciona una metodología para la definición de planes de medición de software observando aspectos organizacionales. Esta metodología podría ser adaptada y extendida con los aspectos específicos de CD a los que antes se hacía referencia para cubrir el hueco existente en el campo de CD. La principal contribución de este artículo al campo de la CD es **MEPLAMECAL,** una metodología basada en la propuesta por ISO/IEC 15939 para la elaboración de planes de medición de CD.

La definición de la metodología implica no sólo la identificación de las actividades y tareas correspondientes, sino también la identificación de los productos de entrada y salida para cada una de ellas. Además, se identifican, a partir de las utilizadas en el campo de Ingeniería del Software, algunas técnicas y herramientas útiles para transformar las entradas en salidas. Pensamos que esta estructura facilitará la incorporación de los aspectos de calidad de datos a las buenas prácticas habituales de calidad del software.

El resto del artículo se estructura de la siguiente manera: la sección 2 presenta un breve repaso al estado del arte de la medición de CD, incluyendo una sinopsis del DQMIM. La sección 3 describe la metodología en sí misma. La sección 4 perfila algunas conclusiones y el trabajo futuro; por último, los agradecimientos y referencias se pueden encontrar al final del artículo.

2 Factores que influyen en la Medición de la Calidad de Datos

Hay un gran número de definiciones para el concepto de CD [2]. De cualquier forma, la mayoría de los autores coinciden en que un dato es de calidad cuando es válido para el propósito para el que un usuario de esos datos quiere utilizarlo ("*adecuación al uso*" [2, 34]). De aquí en adelante, para ganar generalidad en las explicaciones, se ampliará el término "**usuario**" mediante el término

"**implicado**", refiriéndose a *cualquier agente (persona, sistema o proceso) involucrado en el uso de los datos* [37].

La definición anterior de calidad, basada en la adecuación al uso, tiene dos componentes importantes: percepción multidimensional de la calidad y dependencia de contexto. Entorno a estos dos componentes, los implicados deben tener en cuenta los factores a los que se hacía referencia en la introducción de cara a la definición de las medidas de CD.

En cualquier caso, los implicados deben primero identificar las razones por las que quieren medir el nivel de CD de los datos usados por alguno de los procesos de la organización. De acuerdo al DQMIM, se puede decir que los implicados podrían querer satisfacer determinadas **Necesidades de Información**.

Una de las estrategias más usadas para tratar el estudio de la percepción multidimensional de la CD es descomponerla en características más pequeñas, igual que ISO/IEC 9126 [22] hace para el software. Estas características se denominan Dimensiones de CD [2, 24] aunque DQMIM propone el término "**concepto medible**" en su lugar.

La identificación y definición de estos conceptos medibles de CD es todavía un reto para la comunidad de CD debido a su dependencia del contexto. Una definición adecuada permitiría a los investigadores y expertos entender (y, por tanto, medir) la CD, evitando la ambigüedad en las definiciones, términos y conceptos utilizados. Algunos trabajos, como [2, 24, 36] son de obligada referencia para consultar el significado de los conceptos medibles de CD más usados. Para cada escenario en el que se necesita medir la CD de los datos usados, se deben elegir los conceptos medibles más adecuados. El conjunto de conceptos medibles de CD escogidos es conocido como **modelo de CD**. Existen numerosos ejemplos en las referencias bibliográficas de modelos de CD para distintos entornos: salud y asistencia sanitaria [1], militar [5], sistemas de apoyo a la decisión [18], o web [9, 14], por citar algunos. Es importante reseñar que ISO está trabajando actualmente en el estándar ISO/IEC 25012 [21], una parte de la familia de estándares SQUARE que propondrá un modelo de CD para SI. De cualquier forma, y a la espera de que el estándar esté terminado, la clasificación propuesta por [34] es la más utilizada (ver Tabla 1).

Tabla 1. Modelo genérico de CD por [34]

Categoría	Conceptos medibles de CD	Descripción
CD intrínseca	Precisión (*Accuracy*), Objetividad (*Objectivity*), Credibilidad (*Believability*), Reputación (*Reputation*)	Calidad que tienen los datos por sí mismos
CD de accesibilidad	Accesibilidad (*Accessibility*), Seguridad de Acceso (*Access security*)	Suministran significado sobre la facilidad de acceso a los datos
CD contextual	Relevancia (*Relevancy*), Valor añadido (*Value-Added*), Oportunidad (*Timeliness*), Completitud (*Completeness*), Cantidad de datos (*Amount of data*)	Tratan con el uso de los datos en un contexto
CD representacional	Interpretabilidad (*Interpretability*), Facilidad de comprensión (*Ease of understanding*), Representación concisa (*Concise Representation*), Representación consistente (*Consistent representation*)	Características de la representación de los datos que los hacen útiles

Cualquier implicado, sea cual sea su rol, necesitará determinar el grado de bondad de un dato con respecto a los conceptos medibles más adecuados para la tarea que se esté realizando. Esta medida dependerá del uso previsto para los datos y de la naturaleza de los conceptos medibles de CD, que determinan el método o la función de medición [23]. Las típicas medidas de CD tienen una escala de ratio con valores suministrados por la función de medición como, por ejemplo, la fórmula propuesta por [2, 24] y que se muestra en (1):

$$CD_{Medida} = 1 - \frac{NúmeroDeUnidadesDeDatosQueNoSatisfacenUnCriterio}{NúmeroTotalDeUnidadesDeDatos} \quad (1)$$

En la fórmula (1), *UnidadesDeDatos* se refiere a las instancias de los atributos medibles. Además hay dos medidas base:

- *NúmeroTotalDeUnidadesDeDatos*, cuyos valores puede ser calculados de forma objetiva usando como método de medición el conteo de las instancias de los atributos medibles de las entidades cuya CD está siendo evaluada, y
- *NúmeroDeUnidadesDeDatosQueNoSatisfacenUnCriterio*, cuyos valores también pueden ser calculados de una forma objetiva contando el número de "falsos" obtenido después de comprobar si las unidades de datos satisfacen el criterio.

En algunas ocasiones, para asignar valores a la medición de un concepto medible y poder aplicar el criterio, es preciso completar mediante metadatos el significado de una unidad de datos de acuerdo a ese concepto medible. Para tomar una decisión, se necesita además de un valor para el metadato, una regla que usando los metadatos describa el criterio de aceptación como, por ejemplo, la pertenencia del valor del metadato a un dominio determinado. [28] identifica como fuentes de valores para metadatos a los propios implicados (se considera que normalmente proporcionaría un valor subjetivo), al proceso de producción de la información o incluso al mismo almacén de datos.

Para hacer repetibles los procesos de medición, es necesario que el valor del metadato quede adjuntado al dato que completa en el modelo de datos del SI. En [38], se propone una solución para el modelo relacional mediante el etiquetado de los datos como si fueran atributos relacionales. [7] propone un esquema XML llamado DQXSD que permite añadir etiquetas a ficheros XML, [8] incluso propone el uso de Tecnologías Semánticas para realizar esta adjuntado, habilitando así a las aplicaciones Web el procesamiento de los aspectos referidos a la medición de CD.

Otro factor importante que debe ser tenido en cuenta, es la posibilidad de limitar el número de unidades de datos en las que el criterio va a ser aplicado [12] de cara a minimizar el esfuerzo computacional invertido en la medición y que consumiría recursos no dedicados al procesamiento de los datos. Así, para mejorar el rendimiento del SI puede ser necesario seleccionar un conjunto representativo de unidades de datos. El número de unidades de datos y su distribución dependerá de la naturaleza de las necesidades de información.

Para no interferir en el proceso de medición, es necesario fijar en el ciclo de vida cuál es el mejor momento o el mejor intervalo de tiempo para ejecutar el proceso de medición sobre los datos [31].

Como se ha visto a lo largo de este apartado, gestionar todos estos factores puede resultar complicado. Por esto, comprendemos la necesidad de alguna guía para la definición de planes de medición de CD que tenga en cuenta estos factores y su influencia en el proceso de planificación de la medida. A continuación se presenta MEPLAMECAL.

3 Metodología MEPLAMECAL.

El objetivo de esta sección es describir de forma resumida los puntos más interesantes de MEPLAMECAL. La metodología se compone de dos actividades, descritas más adelante, divididas a su vez en tareas. Para cada una de las tareas propuestas, se enumeran tanto los principales productos de entrada como los productos de salida esperados. También se propone el uso de técnicas y herramientas (procedentes en su mayoría del campo de la Ingeniería del Software) para obtener los productos de salida, aunque la elección depende de las preferencias de cada organización. Con el objetivo de facilitar el uso de la metodología, también se identifican los implicados que deberían participar en la ejecución de cada tarea. Las actividades se etiquetan con COM y con EPM, y para las tareas se añade un número que indica el orden en el que deberían ser realizadas. En la Fig. 1 se muestra un diagrama que resume las actividades y tareas de la metodología, junto con su orden de ejecución y los implicados en cada una de ellas.

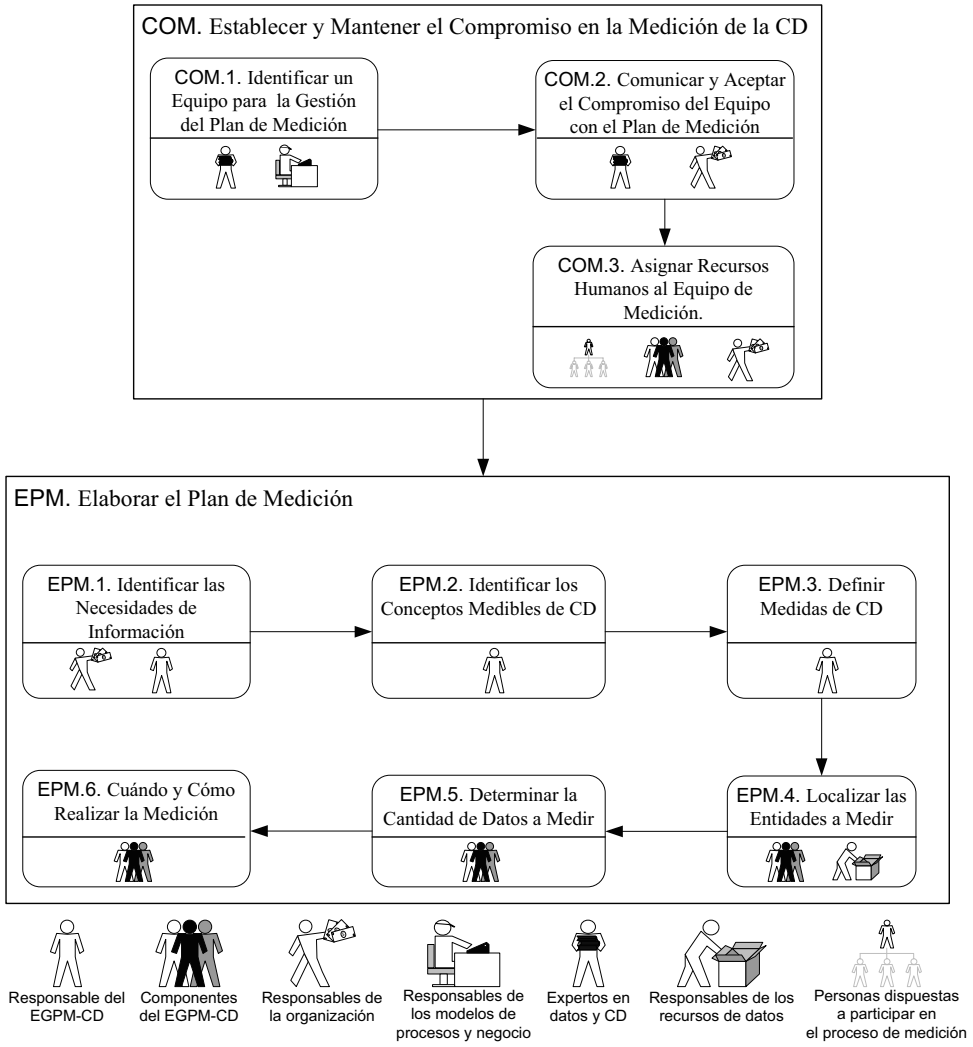

Fig. 1. Diagrama de la metodología MEPLAMECAL.

3.1 COM. Establecer y Mantener el Compromiso en la Medición de la CD.

Los expertos en calidad coinciden en la necesidad de delegar la responsabilidad de la gestión de la calidad a un grupo de gente comprometida de la organización [11]. Para dar soporte a esta necesidad, uno de los objetivos de la metodología es esbozar un equipo multidisciplinar de trabajadores, que puede responsabilizarse de la realización del plan de medición de la CD. Para satisfacer el objetivo propuesto en esta actividad, se deben ejecutar las siguientes tareas.

COM.1. Identificar un Equipo para la Gestión del Plan de Medición.
Este equipo será el responsable de coordinar los esfuerzos y recursos disponibles de la organización de cara a la planificación de la medición de la DQ. El **Equipo de Gestión del Plan de Medición de la CD (EGPM-CD)** debería estar compuesto por roles con responsabilidad directa sobre los datos y su integridad, de forma que se pueda determinar quién está usando los

datos y para qué propósito, con el fin de valorar la naturaleza y alcance de cualquier deficiencia que pueda existir, así como evaluar el impacto que los problemas relacionados con la CD puedan ocasionar. En consecuencia, si se decide externalizar el proceso de medición, el EGPM-CD no debería estar formado exclusivamente por personas externas a la organización sino que sería necesaria la cooperación entre componentes externos e internos para la elaboración del plan.

Dicho equipo debe ser multidisciplinar, cohesionado, sus miembros deben complementarse en conocimientos y cualidades, y tener capacidad de crítica a su propio trabajo. Además deben disponer de un entorno operativo donde se identifique un método de trabajo con formas efectivas de comunicación, así como contar con un catálogo de técnicas y herramientas útiles para cada tarea. La Tabla 2 muestra los principales artefactos para esta actividad.

Tabla 2. Artefactos para COM.1.

Productos	Entrada	- Lista de personas que gestionan o trabajan con los datos de la organización.
	Salida	- Propuesta de Equipo de Gestión del Plan de medición (EGPM-CD).
Herramientas y técnicas		- Entrevistas - Sesiones de trabajo.
Implicados		- Expertos en datos y CD. - Responsables de los modelos de procesos y negocio de la organización.

COM.2. Comunicar y Aceptar el Compromiso del Equipo con el Plan de Medición.
Una vez identificado el EGPM-CD, se debe llevar a cabo una segunda tarea de asignación de personas a los roles componentes del EGPM-CD. Sería interesante llevar a cabo entrevistas y sesiones de trabajo con los candidatos para comprobar su disposición, adecuación y, llegado el caso, su compromiso en la elaboración del plan de medición. La Tabla 3 da más detalles acerca de los artefactos de esta actividad.

Tabla 3. Artefactos para COM.2.

Productos	Entrada	- Propuesta de Equipo de Gestión del Plan de Medición (EGPM-CD). - Organigrama de la organización.
	Salida	- Compromiso de los componentes de la organización para participar en la elaboración del plan de medición. - Documentos de carga de trabajo.
Técnicas y herramientas		- Entrevistas. - Sesiones de trabajo. - Compromiso.
Implicados		- Expertos en datos y CD. - Responsables de la organización.

COM.3. Asignar Recursos Humanos al Equipo de Medición.
Tras conseguir el compromiso de los componentes del EGPM-CD propuesto, hay que estudiar su disponibilidad y restricciones temporales. Para ello, se deberían realizar entrevistas con los responsables de la organización para identificar posibles solapamientos con otros proyectos de la organización que pudieran malograr la ejecución del plan de medición. Algunas herramientas útiles para conseguir este objetivo son los diagramas de Gantt o PERT. Por otra parte, se deberían llevar a cabo sesiones de trabajo y entrevistas para detectar y clasificar habilidades individuales de los componentes del EGPM-CD y, dependiendo de ellas, asignar a cada participante las tareas en las que sería más útil. La Tabla 4 muestra los artefactos correspondientes.

Tabla 4. Artefactos para COM.3.

Productos	Entrada	- Componentes del EGPM-CD.
	Salida	- Composición definitiva del EGPM-CD.
Técnicas y herramientas		- Sesiones de trabajo. - Entrevistas. - Herramientas de planificación temporal como los diagramas de Gantt o PERT. - Estimaciones. - Análisis de coste/beneficio.
Implicados		- Responsables de la organización. - Componentes del EGPM-CD. - Personas dispuestas a participar en el proceso de medición.

3.2 EPM. Elaborar el Plan de Medición.

El principal objetivo de esta segunda actividad es esbozar el Plan de Medición de CD teniendo en cuenta todos los factores mencionados en la sección 2. Para documentar el plan se usarán los términos proporcionados por la terminología DQMIM. Como resultado se obtendrá un documento con el plan de medición de CD. Para conseguir este objetivo, se propone la realización de las siguientes tareas.

EPM.1. Identificar las Necesidades de Información.
Para cada uno de los escenarios donde sea necesaria la medición, es esencial recoger y comprender los requisitos de calidad que deben satisfacer los datos y comprobar que el funcionamiento actual del SI satisface dichos requisitos [17]. Si se encontraran no-conformidades (niveles inadecuados de CD), el EGPM-CD debe profundizar en las causas que las provocaron para arreglarlas o, al menos, mitigar sus efectos. La clasificación propuesta en [29] recoge causas comunes relacionados con los SI que pueden ser utilizados como guía en la identificación de dichas no-conformidades. Puesto que las necesidades de información no se refieren únicamente a aspectos técnicos sino también a organizativos y de gestión, se deberían tener en cuenta las relaciones de negocio existentes entre los principales usuarios del SI. La Tabla 5 muestra los artefactos para esta tarea.

Tabla 5. Artefactos para EPM.1.

Productos	Entrada	- Especificación de requisitos de CD. - Informes de actividad de los SI de la organización.
	Salida	- Lista de necesidades de información.
Técnicas y herramientas		- Entrevistas. - Sesiones de trabajo.
Implicados		- Responsables de EGPM-CD. - Responsables de la organización.

EPM.2. Identificar los Conceptos Medibles de CD.
Los conceptos medibles de CD son criterios racionales que representan los requisitos de usuario para juzgar la CD. El EGPM-CD se debe encargar de seleccionar aquellos conceptos medibles que mejor satisfagan las necesidades de información. Como base para la realización de esta tarea se pueden tomar de la bibliografía aquellos modelos de CD que mejor se adapten al escenario de medición o incluso, si estuvieran disponibles en la organización, aquellos elaborados a partir de experiencias previas. Sobre estos modelos se aplicarán técnicas como tormentas de ideas, sesiones de trabajo, entrevistas o el método Delphi. También se podría utilizar la metodología de Franch y Carvalho descrita en [15] pero adaptada a nuestro campo de calidad de datos. De todos modos, en nuestra metodología se propone continuar la estrategia propuesta por Strong et al. en [33]. En ella se identifican los obstáculos de CD más comunes y los conceptos medibles relativos a cada uno de

ellos. De esta forma, relacionando los obstáculos que afectan a las necesidades de información definidas anteriormente, se pueden obtener los conceptos medibles relativos a cada necesidad de información.

Finalmente, hay tener en cuenta que puede haber dependencias entre los conceptos medibles de CD que forman el modelo, tal y como [10, 16] han analizado, siendo necesario modelar cómo afectan al proceso de medición. En la Tabla 6, se resumen los principales artefactos para esta tarea.

Tabla 6. Artefactos para EPM.2.

Productos	Entrada	- Necesidades de información. - Especificación de requisitos de CD. - Catálogo de conceptos medibles de CD.
	Salida	- Lista de conceptos medibles de CD relevantes. - Lista de entidades a medir.
Técnicas y herramientas		- Sesiones de trabajo. - GQM. - Catalogación. - Lista de obstáculos de Strong et al. en [32]
Implicados		- Responsables del EGPM-CD.

EPM.3. Definir Medidas de CD.

Una vez que los conceptos medibles de CD han sido identificados, es el momento de definir las medidas propiamente dichas, correspondientes a los atributos medibles de las entidades que contienen datos. Tres de las metodologías genéricas más importantes para definir medidas son IEEE 1061, ISO/IEC 15939 y "*Goal-Question-Metric*" (GQM).

De acuerdo al modelo propuesto en DQMIM, las medidas de CD pueden ser de uno de los siguientes tipos: medidas base, medidas derivadas o indicadores. Es importante resaltar que la definición de un plan de medición puede provocar cambios tanto en el modelo de procesos o modelo de datos del SI que contiene los datos cuya calidad se pretende medir. Un ejemplo de este caso es la inserción de metadatos tal y como se describió en la sección 2.

Para cada concepto medible de CD, se debe especificar un procedimiento de medición dependiendo del tipo de medida. Esto lleva aparejado la identificación de una unidad de medida, de una escala y de los correspondientes elementos para completar la definición de la medida. Por ejemplo, para un indicador se debe añadir un criterio de decisión y un modelo de análisis.

Dada la complejidad de la situación, es importante remarcar que el EGPM-CD debe ser multidisciplinar y con habilidades suficientes como para afrontar esta actividad, que puede ser considerada como una de las más importantes de la metodología presentada. La Tabla 7 muestra los principales artefactos para esta tarea.

Tabla 7. Artefactos para EPM.3.

Productos	Entrada	- Necesidades de información. - Conceptos medibles relevantes. - Modelo de datos y procesos.
	Salida	- Lista de medidas a aplicar para cada concepto medible.
Técnicas y herramientas		- GQM.
Implicados		- Responsables del EGPM-CD.

EPM.4. Localizar las Entidades a Medir.

Las entidades a medir se pueden encontrar en almacenes de datos de distinta naturaleza como, por ejemplo, bases de datos relacionales o semi-estructuradas, ficheros de acceso secuencial, documentos XML u hojas de cálculo.

Para planificar la medición, es preciso localizar la ubicación de las entidades cuyo nivel de CD va a ser medido. Estas entidades tienen atributos medibles que deben ser inspeccionados a la hora

de trazar el plan de medición. Algunos ejemplos de entidades pueden ser esquemas de datos, valores de datos, dominios de datos, reglas de negocio, o interfaces de usuario [7].

Una de las principales técnicas para localizar estas entidades son las sesiones de trabajo con los responsables de los recursos de datos de la organización. En la Tabla 8, se resumen los principales artefactos para esta tarea.

Tabla 8. Artefactos para EPM.4.

Productos	Entrada	- Lista de entidades a medir. - Modelo de datos de la organización.
	Salida	- Lista de la ubicación de los repositorios de datos a medir.
Técnicas y herramientas		- Inspecciones. - Sesiones de trabajo.
Implicados		- Componentes del EGPM-CD. - Responsables de los recursos de datos.

EPM.5. Determinar la Cantidad de Datos a Medir

Dependiendo del propósito de la medición o la necesidad de no sacrificar el rendimiento del SI, puede resultar necesario delimitar el número de entidades que se deben tener en cuenta para la medición. En tal caso, se debe extraer una muestra estadísticamente representativa de todo el conjunto de entidades y luego extrapolar los resultados. Los parámetros de la muestra (ratio aceptable de datos no válidos, tamaño de la muestra, tipo de muestreo, máximo y mínimo valor para la aceptación o rechazo) pueden ser calculados de acuerdo a estándares como ISO 2859 o UNE 66020, siempre y cuando las entidades sobre las que se va a medir verifiquen las condiciones y limitaciones impuestas por estos estándares. En la Tabla 9 se pueden ver los artefactos para esta tarea.

Tabla 9. Artefactos para EPM.5.

Productos	Entrada	- Necesidades de información. - Repositorios de datos a medir. - Coste computacional necesario para medir la colección de datos completa. - Relación esfuerzo-coste para la medición de la colección de datos completa.
	Salida	- Estudio de viabilidad de la medición de la colección de datos completa. - Cantidad de datos a medir para cada entidad.
Técnicas y herramientas		- UNE-EN-ISO66020 / ISO 2859.
Implicados		- Componentes del EGPM-CD.

EPM.6. Cuándo y Cómo Realizar la Medición.

La medición puede no ser una actividad puntual sino que su realización requiera un tiempo, o simplemente puede ocurrir que alguien esté interesado en estudiar la evolución temporal del nivel de CD de una entidad [31]. También es fundamental considerar la asignación temporal que se hizo durante la tarea COM.3 para evitar colisiones temporales entre los componentes del EGPM-CD que desarrollen otras labores dentro de la organización aparte de su intervención en la planificación de la medición.

Por estas razones, es necesario realizar una planificación temporal de la medición de modo que se puedan obtener los valores más significativos, y no se entorpezca el trabajo realizado por los implicados. Algunas herramientas que se pueden utilizar para la planificación temporal son los diagramas de Gantt o incluso, si se pretendiese localizar en determinados puntos del proceso de negocio, se podría utilizar alguna notación de modelado de procesos de negocio, como BPMN[30] o IPMAP, notación específica de calidad de datos desarrollada por Shankaranarayan et al. [32].

Una idea muy interesante sería automatizar los procedimientos de medición de cara a ahorrar recursos necesarios en la ejecución del plan de medición y obtener resultados más fiables.

Para poder interpretar adecuadamente los resultados, es importante describir apropiadamente la forma en que los resultados de la medición serán comunicados como parte del Plan de Medición [20, 35].

En la Tabla 10 se recogen los principales artefactos para esta tarea.

Tabla 10. Artefactos para EPM.6.

Productos	Entrada	- Lista de medidas a aplicar. - Lista de entidades a medir. - Cantidad de datos a medir para cada entidad. - Planificación temporal de la actividad de la organización. - Ciclo de vida de los datos.
	Salida	- Plan de medición.
Técnicas y herramientas		- Diagramas de Gantt. - BPMN, IPMAP, Diagramas de actividad UML. - Sesiones de trabajo
Implicados		- Componentes del EGPM-CD.

4 Conclusiones y Trabajo Futuro

La medición es una actividad clave en cualquier iniciativa de gestión de la calidad. Dado que las organizaciones han empezado a darse cuenta de que algunos de los problemas que sufren son debidos a niveles inadecuados en la calidad de los datos, han comenzado a dedicar esfuerzos y recursos orientados a gestionar la calidad de los datos usados en sus procesos de negocio.

En este sentido, los investigadores en el campo de la CD están comprometidos con el desarrollo de artefactos que ayuden a las organizaciones a medir su CD. Como el campo de la CD es todavía muy joven, necesita apoyarse en otros más consolidados, como el de la Ingeniería del Software, de modo que se puedan basar los esfuerzos de investigación en fundamentos sólidos, como los estándares existentes para la medición y la calidad del software. Hemos elegido uno de estos estándares, ISO/IEC 15939, como la base para la metodología MEPLAMECAL, presentada en este trabajo. Nuestra propuesta pretende llenar un vacío existente y además complementar a los trabajos de evaluación de CD ya existentes en la bibliografía.

Entendemos que la diferencia fundamental entre medición y evaluación consiste en el enfoque específico de ésta última de determinar la validez y utilidad de los datos dentro de un contexto; mientras que la medición está sólo encaminada a la obtención de valores que se utilizarán en la evaluación, sin dar soporte a juicio alguno. A pesar de que pueda parecer que esta simplificación resta relevancia a nuestra propuesta, no existe en la literatura ninguna iniciativa similar con la suficiente generalidad, y que además permita tener en cuenta las características especiales de los datos así como los aspectos inherentes a la medición de su calidad dentro del contexto organizacional.

La principal contribución de este artículo no es la metodología en sí, sino la ventaja de poder utilizarla como guía eficiente que tiene en cuenta los aspectos anteriormente citados en la planificación de la medición de la CD de las entidades organizacionales que contienen los datos que intervienen en los procesos de negocio.

MEPLAMECAL consiste en dos actividades con sus correspondientes tareas. Para facilitar el uso de la metodología, se han identificado los artefactos para cada una de estas tareas. Dado que la metodología ha sido adaptada a partir del estándar ISO/IEC 15939, muy conocido en el campo de la medición de software, cualquier profesional familiarizado podría aplicar fácilmente la metodología en su propio contexto e ir introduciendo en su catálogo de buenas prácticas los conceptos de calidad de datos.

Finalmente, es importante resaltar que para conseguir que el proceso de medición sea lo más efectivo posible es necesario que no se vea como una actividad puntual o un proceso aislado, sino como un proceso más que se integra dentro de la actividad diaria de la organización.

Somos conscientes de la importancia de obtener medidas de CD repetibles. En consecuencia, nuestra línea de trabajo actual se centra en dos aspectos: por un lado, validar la metodología mediante su aplicación a casos prácticos, con el objetivo de identificar las necesidades de información que podrían aceptar medidas con procedimientos de medición automatizables; y por otro lado, estamos desarrollando un conjunto de herramientas para la automatización del propio proceso de elaboración de Planes de Medición de CD mediante el uso de Tecnologías Semánticas.

Agradecimientos

Esta investigación es parte de los proyectos ESFINGE (TIN2006-15175-C05-05), CALIPSO (TIN 2005-24055-E), ambos apoyados por el Ministerio de Educación y Ciencia, y HERMES (TSI-020100-2008-155) apoyado por el Ministerio de Industria, Turismo y Comercio.

Referencias

1. Al-Hakim, L., *Procedure for Mapping Information Flow: A case of Surgery Management Process*, in *Information Quality Management: Theory and Applications*, L. Al-Hakim, Editor. 2007, Idea Group Publishing: Hershey, PA, USA. p. 168-188.
2. Batini, C. and M. Scannapieco, *Data Quality: Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. 2006, Berlin: Springer-Verlag Berlin Heidelberg.
3. Bobrowski, M., M. Marrá, and D. Yankelevich. *A Software Engineering View of Data Quality*. in *Second International Software Quality in Europe*. 1998. Brussels, Belgium.
4. Burgess, M.S.E., W.A. Gray, and N.J. Fiddian. *Quality Measures and the Information Consumer*. in *Ninth International Conference on Information Quality (ICIQ'04)*. 2004. MIT, Cambridge, MA, USA.
5. Burzynski, T. *Establishing the Environment for Implementation of a Data Quality Management Culture in the Military Health System*. in *Third International Conference on Information Quality (ICIQ'98)*. 1998. MIT, Cambridge, MA, USA.
6. Caballero, I., Ó. Gómez, and M. Piattini. *Getting Better Information Quality by Assessing and Improving Information Quality Management*. in *Ninth International Conference on Information Quality (ICIQ'04)*. 2004. MIT, Cambridge, MA, USA.
7. Caballero, I., et al. *A Data Quality Measurement Information Model based on ISO/IEC 15939*. in *12th International Conference on Information Quality*. 2007. MIT, Cambridge, MA.
8. Caballero, I., et al. *DQRDFS:Towards a Semantic Web Enhanced with Data Quality*. in *Web Information Systems and Technologies*. 2008. Funchal, Madeira, Portugal.
9. Caro, A., et al., *A proposal for a set of attributes relevant for Web Portal Data Quality*. Software Quality Journal, 2008.
10. DeAmicis, F., D. Barone, and C. Batini. *An Analytical Framework to analyze Dependencies among data Quality Dimensions*. in *ICIQ'06*. 2006. MIT, Cambridge, MA, USA.
11. Deming, W.E., *Out of Crisis*. 1986, Cambridge: MA: MIT Center for Advanced Engineering Study.
12. English, L., *Improving Data Warehouse and Business Information Quality: Methods for reducing costs and increasing Profits*. 1999, New York, NY, USA: Willey & Sons.
13. Eppler, M. and M. Helfert. *A Classification and Analysis of Data Quality Costs*. in *International Conference on Information Quality*. 2004. MIT, Cambridge, MA, USA.
14. Eppler, M. and P. Muenzenmayer. *Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology*. in *Proceeding of the Seventh International Conference on Information Quality*. 2002.
15. Franch, X. and J.P. Carvallo, *Using Quality Models in Software Package Selection*. IEEE Software., 2003. **20**(1): p. 34-41.
16. Ge, M. and M. Helfert. *A Review of Information Quality Research*. in *Interantional Conference on Information Quality*. 2007. MIT, Cambridge, MA, USA.

17. Gebauer, M., P. Caspers, and N. Weigel. *Reproducible Measurement of Data Quality Field.* in *Tenth International Conference on Information Quality (ICIQ'05).* 2005. MIT, Cambridge, MA, USA.
18. Gendron, M. and M.J. D'Onofrio. *Formulation of a Decision Support Model Using Quality Attributes*. in *Seventh International Conference on Information Quality (ICIQ'02).* 2002. MIT, Cambridge, MA, USA.
19. Gustavsson, M. *Information Quality Measurement*. in *International Conference on Information Quality.* 2006. MIT, Cambdrige, MA, USA.
20. Humphrey, W.S., *Managing the Software Process*. The SEI Series in Software Engineering. 1989, Pittsburgh, PA Addison-Wesley.
21. ISO-25012, *ISO/IEC 25012: Software Engineering - Software Quality Requirements and Evaluation (SQuaRE) - Data Quality Model (Draft)*. 2006.
22. ISO/IEC, *ISO/IEC 9126. Software Engineering-Product Quality. Parts 1 to 4*. 2001: International Organization for Standardization/International Electrotechnical Commission.
23. ISO/IEC, *ISO/IEC 15939. Information Technology - Software Measurement Process*. 2000.
24. Lee, Y.W., et al., *Journey to Data Quality.* 2006, Cambridge, MA, USA: Massachussets Institute of Technology.
25. Lee, Y.W., et al., *AIMQ: A Methodology for Information Quality Assessment*. Information and Management, 2002. **40**(2): p. 133-146.
26. Levis, M., M. Helfert, and M. Brady. *Information Quality Management: Review of an Evolving Research Area*. in *ICIQ'07*. 2007. MIT, Cambridge, MA, USA.
27. Loshin, D., *Enterprises Knowledgement Management: The Data Quality Approach.* 2001, San Francisco, CA, USA: Morgan Kauffman.
28. Naumann, F. and C. Rolker. *Assessment Methods for Information Quality Criteria.* in *Fifth International Conference on Information Quality (ICIQ'2000).* 2000. MIT, Cambridge, MA, USA.
29. Oliveira, P., et al. *A Taxonomy of Data Quality Problems*. in *Second International Workshop on Data and Information Quality (in conjunction with CAISE'05).* 2005. Porto, Portugal.
30. OMG, *Business Process Model and Notation 2*. 2008, Object Management Group.
31. Redman, T., *Data Quality: The field guide*. 2000, Boston: Digital Press.
32. Shankaranarayan, G., R.Y. Wang, and M. Ziad. *IP-MAP: Representing the Manufacture of an Information Product*. in *Fifth International Conference on Information Quality (ICIQ'2000).* 2000. MIT, Cambridge, MA, USA.
33. Strong, D., Y. Lee, and R. Wang, *Ten Potholes in the Road to Information Quality*. IEEE Computer, 1997: p. 38-46.
34. Strong, D.M., Y.W. Lee, and R.Y. Wang, *Data Quality in Context*. Communications of the ACM, 1997. **40**(5): p. 103-110.
35. Ukko, J., J. Karhu, and H. Rantanen, *How to communicate measurement information successfully in small and medium-sized entreprises: a regression model*. International Journal of Information Quality, 2007. **1**(1): p. 41-59.
36. Wang, R., et al., eds. *Information Quality.* Advances in Management Information Systems, ed. V. Zwass. 2005, M.E. Sharpe: Saddle River, NJ.
37. Wang, R.Y., *A Product Perspective on Total Data Quality Management*. Communications of the ACM, 1998. **41**(2): p. 58-65.
38. Wang, R.Y., M. Reddy, and H. Kon, *Towards quality data: An attribute-based approach*. Journal of Decision Support Systems, 1995. **13**(3-4): p. 349-372.