Actas del Taller de Trabajo Zoco'09 / JISBD

Integración de Aplicaciones e Información Empresarial
XIV Jornadas de Ingeniería del Software y Bases de Datos
San Sebastián, 8 de septiembre de 2009

http://www.tdg-seville.info/cfp/zoco/

Organizadores

José L. Álvarez, José L. Arjona, Iñaki Fernández de Viana (Universidad de Huelva)
Rafael Corchuelo, David Ruiz, Carlos Rivero, Hassan A. Sleiman, Inmaculada Hernández (Universidad de Sevilla)

Ponentes

Eduardo Martín Rojo y Vicente Luque Centeno (Universidad Carlos III de Madrid) □ Hassan A. Sleiman (Universidad de Sevilla) □ Patricia Jiménez (Universidad de Huelva) □ Carlos G. Figuerola, José Luis Alonso Berrocal y Angel Zazo Rodríguez (Universidad de Salamanca) □ Inma Hernández (Universidad de Sevilla) □ Paula Montoto, Alberto Pan, Juan Raposo, Fernando Bellas y Javier López (Universidade da Coruña) □ M. Mercedes Martínez-González, Beatriz Pérez-León (Universidad de Valladolid) y M. Luisa Alvite-Díez (Universidad de León) □ Ana Flores Cuadrado (Telefónica Investigación y Desarrollo), Eduardo Villoslada de la Torre (Telefónica Investigación y Desarrollo, Universidad de Valladolid) y Alberto Peláez Gutiérrez (Telefónica Soluciones) □ Ismael Navas-Delgado, Amine Kerzazi y José F. Aldana-Montes (Universidad de Málaga) □ María Pérez, Ismael Sanz, María José Aramburu y Rafael Berlanga (Universitat Jaume I de Castelló) □ Ismael Caballero, Mª Ángeles Moraga y Coral Calero (Universidad de Castilla–La Mancha) □ Juan A. Fraile, Javier Bajo (Universidad Pontificia de Salamanca) y Juan M. Corchado (Universidad de Salamanca) □ Francisco J. Garijo (Telefónica I+D), Juan Pavón, Carlos Rodríguez y Damiano Spina (Universidad Complutense de Madrid)

Agradecimiento

Financiación Proyecto IntegraWeb (TIN2007-64119, P07-TIC-02602, P08-TIC-4100)

Índice

Integración de Aspectos de Calidad de Datos en Sistemas de Información

Ismael Caballero, Mª Ángeles Moraga, Coral Calero

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

Abstract. Para desarrollar con éxito una tarea es imprescindible que los usuarios de información puedan disponer de datos con los niveles adecuados de calidad. Para tomar una decisión sobre si usar o no los documentos, necesitan tener una evaluación de esos niveles. Para evaluarlos, es necesario disponer de un modelo de calidad de datos. Como quiera que la evaluación depende del contexto, y unos mismos datos deben estar disponibles para varios contextos, entonces toda la infraestructura de evaluación debe estar disponible para ser accedida desde distintas aplicaciones. Esto requiere que los datos estén adecuadamente descritos para que puedan ser entendidos. Este artículo plantea algunos de los retos de investigación que estamos afrontando, como por ejemplo la posibilidad de aplicar los principios de LinkedData para anotaciones de calidad de datos, o como aplicar los resultados que vayamos obteniendo a los datos almacenandos en bases de datos heredadas, de modo que puedan ser utilizados en entornos distintos para los que fueron plantearon, de forma que puedan seguir siendo utilizados.

Palabras Clave: Calidad de Datos, Dimensiones de Calidad de Datos, Web Semántica, Integración, Evaluación de Calidad de Datos, Linked Data.

1. Introducción

Los usuarios necesitan datos para desarrollar tareas asociadas a su trabajo. Estos datos están recogidos o agrupados en documentos, o que pueden ser obtenidos consultando bases de datos. Habitualmente estos datos son reutilizados en diferentes tareas por diferentes usuarios desde distintas aplicaciones. Por ejemplo, una historia clínica electrónica de un paciente puede ser usada (leída) por un médico para diagnosticar una enfermedad, o puede ser usada (actualizada) por un analista de un laboratorio para incluir los resultados de un análisis de sangre, o puede ser usada (leída y actualizada) por un enfermero para consultar el tratamiento que se le debe dar a un enfermo durante su hospitalización y apuntar los datos requeridos en el seguimiento. Cada uno de estos tres actores podría estar usando una aplicación diferente, pero todos ellos están trabajando sobre los datos relativos a un paciente, datos que en este

caso están agrupados en un documento conocido como "historia clínica electrónica de un paciente", y que va evolucionando con la vida del paciente.

Para tener éxito en sus respectivas tareas, estos actores necesitan que los datos contenidos en los documentos tengan niveles adecuados de calidad: precisamente en el ejemplo propuesto, cualquier error derivado de una decisión errónea puede ser crítico para el paciente. Se dice que los datos tienen calidad si sirven para el propósito (tarea) para el que se pretenden usar [1]. Esta definición de calidad de datos basada en *fitness for use* pretende ser más general que la percepción habitual de que los datos son de calidad si no tienen defectos. Esta definición conlleva dos implicaciones importantes: por un lado está la percepción multidimensional del problema (es necesario evaluar la calidad de los datos contenidos en un documento usando distintos criterios, a los que como se verá en la sección 2, se les denominan dimensiones de calidad de datos), y por otro lado se incluye el carácter subjetivo de la calidad de datos (cada uno de los actores anteriores, médicos, analistas de laboratorio, enfermeros, ... puede tener una percepción diferente del nivel de calidad, incluso para la misma dimensión de calidad de datos). Como puede entenderse, estas dos implicaciones son más difíciles de gestionar que la percepción de que los datos tengan "*cero defectos*".

De estas dos implicaciones, se puede deducir que estimar cada una de las dimensiones de calidad requeridas no es tarea fácil. Una forma de realizar esta estimación es mediante la definición de medidas capaces de cuantificar la calidad de cada dimensión. Para poder definir dichas medidas es preciso entender el contexto del usuario y lo que puede entender por el hecho de que los datos se ajusten a su uso. En algunas ocasiones y para ciertas dimensiones de calidad de datos, puede ser posible establecer medidas objetivas (y repetibles) sobre los propios datos de manera independiente del contexto, mientas que en otras muchas, es necesario disponer de datos adicionales (normalmente denominados "*metadatos*" [2]) que de alguna forman representan el contexto de los usuarios. De hecho, podría decirse que estos datos adicionales permiten completar el significado del dato en el sentido marcado por la dimensión de calidad elegida. Así por ejemplo, para determinar el grado de *actualidad* (véase sección 2) de un valor en la bolsa es necesario adjuntar al dato correspondiente al valor del título una fecha que indique cuándo ha sido proporcionado; o por ejemplo, para determinar si una opinión en un foro merece ser tenida en cuenta (grado de *reputación*), un usuario debería poderse conocer quién dio esa opinión (seguramente el foro tenga establecido un sistema de clasificación de reputación de usuarios basado en la antigüedad de los usuarios, o en el número de opiniones proporcionada que son aceptadas como útiles por otros usuarios, ...).

En cualquier caso se hace imprescindible adjuntar y hacer persistentes dichos datos adicionales junto a los datos utilizados por los usuarios, de modo que las distintas aplicaciones puedan disponer de ellos para establecer sus propios métodos de medición. Es importante resaltar que este dato adicional no tiene por qué ser el valor de la medida en sí, o una anotación directa de calidad como proponen Price y Shanks en [3], sino que generalizando, podemos considerar que es un elemento que permitirá calcular la medida de calidad de datos para una determinada dimensión. La bibliografía recoge algunas formas de mantener estos vínculos, tales como la propuesta de Wang y Madnick para el modelo relacional en [2] o la de Caballero et al en [4] para documentos XML.

Nuestra motivación inicial para esta investigación está basada en el hecho de que recientemente, en un proyecto de I+D relacionado con la gestión de historias clínicas electrónicas, nos planteábamos aplicar la propuesta presentada en [2] a una base de datos XML nativa, para optimizar, basándonos en la calidad de datos de los documentos usados, algunos de los procesos que se estaban ejecutando en un sistema de información relacionado con gestión sanitaria, y que ya estaba en producción. La idea era establecer los mecanismos adecuados para medir la calidad de los datos almacenados en dicha base de datos para proporcionar a los usuarios una estimación del nivel de calidad que tenían los datos, de modo que pudieran decidir si usarlos o no en función del nivel de calidad obtenido. Pero pronto nos dimos cuenta que esto implicaba modificar tanto el modelo de procesos como el modelo de datos del SI, pues teníamos que almacenar los datos adicionales y agregar las consultas correspondientes. Tras inspeccionar algunas alternativas, llegamos a la conclusión de que estando el sistema en producción no podíamos hacer las modificaciones correspondientes, al menos las correspondientes al modelo de datos, pues implicaría hacer un mantenimiento adaptativo del resto de la aplicación, extremo que la empresa que ejecutaba el software no se podía permitir. Además surgió otro inconveniente: de hacerlo así, los datos sólo estarían disponibles para una determinada aplicación, y sería muy complicado integrar los resultados en otras aplicaciones, por lo que podía no ser una propuesta suficientemente viable.

Tras inspeccionar los trabajos propuestos por Missier et al en [5] o Stivilia et al en [6] basados en la aplicación de tecnologías semánticas para gestionar aspectos de calidad de datos, llegamos a la conclusión de que la solución pasaba por crear una capa adicional que permitiese gestionar los aspectos relacionados con la medición de la calidad de los datos, y que además permitiera tenerlos disponibles para distintas aplicaciones.

En este artículo presentamos una parte de los avances que hemos ido haciendo al crear tanto la capa, a la que hemos llamado **mapas de agregación,** como los aspectos tecnológicos asociados.

El resto del artículo está estructurado como sigue: en la sección 2 se presenta los conceptos relevantes relacionados con calidad de datos que debemos tener en cuenta para esta gestión. En la sección 3 presentamos los fundamentos de los mapas de agregación, y finalmente, en la sección 4 presentamos algunas conclusiones y líneas de trabajo futuro.

2 Conceptos de Calidad de Datos

Existen numerosas definiciones del concepto de calidad de datos, pero la más aceptada es sin duda aquella basada en la idea de *fitness for use* [7]. Siguiendo esta definición, una de las estrategias más comunes para afrontar el estudio de la calidad de datos (en adelante DQ, de las siglas en inglés Data Quality, por simplificar) consiste en descomponer la percepción global de calidad de datos de un conjunto de datos en una serie de "*subcalidades*" denominadas **dimensiones de DQ** [7], (al estilo de lo que hace la ISO 9126 [8] para el software) y establecer determinados criterios de aceptación para cada una de ellas. Estos criterios de aceptación deben estar definidos

en función de los posibles rangos de los resultados de medición. Así, si tras medir el nivel de calidad según cada una de las dimensiones de DQ del documento, se obtienen resultados fuera de los rangos de aceptación preestablecidos, entonces el documento podría considerarse como "*defectuoso*" y, por tanto, podría considerarse el no usarlo para la tarea, o al menos saber que se puede estar cometiendo un error al usarlo. La situación deseable es que existiese un sistema de medición integrado en el sistema de información que fuera capaz de identificar esos datos o documentos "defectuosos" y recomendar al usuario, al menos, su "no utilización".

Al conjunto de dimensiones de DQ que se utiliza en un contexto determinado para determinar el nivel de DQ de un documento se le denomina **modelo de DQ.** Para cada contexto (entendiéndose como tal el ámbito de una tarea) es necesario identificar aquellas dimensiones que mejor representen los requisitos de DQ de los usuarios. La bibliografía muestra ejemplos de modelos de DQ específicos para ciertos contextos: hospitalarios y de salud [9], o web [10], por nombrar unos cuantos representativos. Así mismo es interesante mencionar que ISO, ha publicado recientemente el estándar ISO/IEC 25012 [11], como parte de la familia SQUARE y que dicho estándar propone un modelo genérico de DQ para Sistemas de Información. En cualquier caso, el modelo que más se viene utilizando es el propuesto por Strong et al. en [1]. Estos autores agrupan las dimensiones DQ en cuatro categorías, que hacen referencia a los puntos de vista desde los cuales es posible evaluar la calidad de los datos (Tabla 1).

Categoría	**Descripción**
Intrínseca de DQ	Con dimensiones como *precisión, objetividad, credibilidad,* y *reputación.* Se refiere a la calidad que tienen los datos por sí mismos;
Accesibilidad de DQ	Con dimensiones como *accesibilidad, seguridad en el acceso.* Están orientadas a determinar si el alcance de los mecanismos de seguridad es suficiente como para poder acceder a los datos en condiciones de poder ser usados adecuadamente
Contextual DQ	Que engloba dimensiones como *relevancia, valor añadido, actualidad, compleción, cantidad apropiada de datos.* Estas dimensiones se refieren al alcance en que los datos pueden ser usados en un contexto determinado.
Representacional de DQ	Que engloba dimensiones como *interpretabilidad, facilidad de compresión, grado de representación concisa, grado de representación consistente.* Estas dimensiones se orientan a determinar si la forma en la que se representan y/o transmiten los datos les hace o no usables.

Tabla 1. Categorías de dimensiones definidas por Strong et al en [1] .

Una definición más completa del significado de estas dimensiones, asumiendo la posibilidad de la dependencia del contexto, puede encontrarse en distintos trabajos en la bibliografía, aunque los más interesantes son, a nuestro juicio, aquellos presentados en [7, 12, 13].

Siguiendo con el proceso de evaluación, el siguiente paso es definir propiamente cómo medir la DQ de los datos para cada una de estas dimensiones. Dado el gran número de propuestas en la bibliografía relacionadas con los aspectos de medición de datos, para una mejor comprensión se recomienda seguir la nomenclatura basada en ISO/IEC 15939 proporcionada por Caballero et al. en [4]. En este trabajo se intenta unificar las diferentes terminologías referidas a la medición de DQ encontradas en la literatura. Además destaca la existencia de tres tipos de medidas: las medidas base (que requieren un método de medición), las derivadas (que requieren una función de cálculo) y los indicadores (que requieren de un modelo de análisis y de un criterio de decisión). En la bibliografía [7, 12], se suele recoger una función de medición basada

en el porcentaje (ratio) de unidades de datos que satisfacen o no un criterio, tal y como está representado en la Figura 1. En base a esta fórmula se suelen definir la mayoría de las funciones de medición de DQ.

$$DQ_{Medida} = 1 - \frac{NúmeroDeUnidadesDeDatosQueNoSatisfacenUnCriterio}{NúmeroTotalDeUnidadesDeDatos}$$

Figura 1. Una Función de Medición de Calidad de Datos.

En la fórmula de la Figura 1, se pueden apreciar dos medidas base: *NúmeroDeUnidadesDeDatosQueSatisfacenUnCriterio* y *NúmeroTotalDeUnidades-DeDatos*. El método de medición de ambas consiste en un conteo del número de unidades. Si para la segunda no hay mayor complicación que contar (por ejemplo, con un *select count(*) from NombreTabla* en SQL), para la primera aparece el problema de expresar, usando la tecnología asociada al sistema de información, si se cumple o no el criterio. Un criterio se puede definir como una **regla de negocio** que define cuándo un dato es válido desde el punto de vista del negocio (tanto de la semántica como de la sintaxis). Según Loshin en [14], esta regla de negocio puede recoger parte del *know-how* empresarial (por ejemplo los sueldos de los trabajadores según sus categorías), la legislación relacionada (como por ejemplo, edades legales para trabajar), u otros aspectos [13, 15, 16]. En cualquier caso, el resultado de decidir si una unidad de datos satisface un criterio puede ser por simplicidad "*Verdadero*" o "*Falso*", aunque en algunas otras ocasiones es posible permitir distintas gradaciones del nivel de cumplimiento ("*Mucho*", "*Normal/Suficiente*", "*Nada*"). Usando la formulación más sencilla, y a fin de obtener un valor para una determinada medida, hay que contar cuántas unidades de datos obtienen un resultado de "*Verdadero*" en la prueba de *no satisface un criterio*. En entornos relacionales este conteo puede ser fácilmente resuelto introduciendo en la instrucción de consulta la cláusula *where*, siempre que sea factible representar así la condición del criterio: por ejemplo contar el número de valores nulos para la dimensión de compleción, o que el valor del dato pertenezca a un dominio definido como parte de la base de datos (en el caso de que el sistema gestor de bases de datos permitiera definir dominios), o bien en el caso de que el dominio se pudiera definir como un conjunto de valores almacenados en otra tabla y relacionados mediante integridad referencial: si está adecuadamente implementado, un intento de inserción de valores que violase el dominio generaría el correspondiente error, como es sabido. En estos casos los criterios pueden ser establecidos objetivamente e independientemente de cualquier otra evidencia.

Como se dijo anteriormente, para evaluar algunas dimensiones de calidad de datos se requiere la emisión de un juicio, individual o colectivo, sobre el valor del dato. Pero muchas veces el dato por sí mismo no aporta suficiente información para emitir dicho juicio, sino que es necesario hacerlo sobre un dato adicional (en [5] son llamados "evidencias") que completa el significado del dato en la dirección de la dimensión que se pretende evaluar. Este dato de juicio puede ser objetivo o subjetivo y se puede realizar sobre datos adicionales proporcionados de manera también objetiva o subjetiva. La Tabla 2, recogida en [4], muestra algunos ejemplos de escenarios donde se emiten juicios objetivos o subjetivos teniendo en cuenta datos adicionales proporcionados.

	Juicio Objetivo	Juicio Subjetivo
[redacted]	El estudio de la *actualidad* de una oferta de trabajo publicada en Internet con una fecha de inicio y otra fecha de final de validez. El portal Web debería proporcionar ambas fechas a los usuarios, de modo que puedan hacer un juicio basándose en evidencias objetivas al comparar la actualidad de la oferta con la fecha actual.	El estudio del *Valor Añadido* de un dato: Alguien interesado en una cámara digital con zoom óptico mira las especificaciones técnicas del producto que busca para saber si el fabricante proporciona además información sobre el sistema de alimentación (pilas o baterías). Si lo encuentra el usuario debe decidir si los datos proporcionados hacen mejor o no la información que tiene sobre la cámara.
[redacted]	En el estudio de la *reputación:* Alguien está interesado en ver una película y busca críticas en internet o en un periódico. Puede tener más ganas de ver la película o perder el interés en ella, si la crítica que encuentra está hecha por alguien que se considera que tiene siempre un buen criterio a la hora de hacer las críticas.	El estudio de la *credibilidad* de una noticia publicada en un periódico y que ha sido proporcionada por una agencia de noticias. Si alguien tiene prejuicios de que las noticias publicadas por ese periódico defienden unas determinadas ideas políticas, en caso de tener ideas distintas podría no creer las noticias publicadas, o en caso contrario creer la mayoría de ellas.

Tabla 2. Ejemplos de tipos de escenarios donde la calidad de los datos se mide en función de datos adicionales proporcionados [4].

De todos modos, se requiere tener tanto el o los datos adicionales necesarios, almacenados y enlazados convenientemente junto con el dato cuya calidad se pretende evaluar y con la correspondiente regla de negocio. Como puede apreciarse, la dificultad está en mantener esta infraestructura, y es incluso más complicado para aplicaciones ya desarrolladas. Se necesita por tanto describir y almacenar adecuadamente la regla de negocio para poder reutilizarla después para otras instancias del proceso de medición, bien dentro de la misma aplicación, bien como parte de otras aplicaciones. En cualquier caso, es patente la necesidad de modificar tanto el modelo de datos como el modelo de negocio de la aplicación. Este paso depende obviamente de la tecnología usada en el desarrollo de la aplicación: sistemas gestores de bases de datos (SGBD) más o menos avanzados cuentan con lenguajes procedimentales embebidos, PLSQL o TransactSQL, para acometer estas tareas, pero no todas las empresas pueden tener o tienen posibilidad de utilizar estos SGBDR. Además nuevos modelos de negocio demandan el uso de otras tecnologías y arquitecturas que permitan el acceso remoto a través de mecanismos bien definidos y asequibles como puede ser http. Precisamente, este va a ser uno de los muros que tenemos que superar en nuestra propuesta.

Otra dificultad añadida es la forma de operar con los resultados obtenidos de los procesos individuales de medición para cada una de las dimensiones: si todos los procesos de medición proporcionan como resultado medidas expresadas en la misma escala numérica (típica y preferiblemente en formato ratio como se describió anteriormente), entonces la evaluación de un dato se hace mediante una media ponderada entre las dimensiones y los resultados de la medición [12, 17], aunque no debe descartarse el uso de etiquetas lingüísticas, que necesitan de operadores específicos para obtener un resultado agregado de todas ellas [18, 19]. Es importante destacar que normalmente, y en aras de la simplicidad, los modelos de calidad de datos dan la misma importancia (peso) a todas las dimensiones de calidad que la

componen. Pero, como Wang y Strong demostraron en [20], no todas estas dimensiones son igualmente importantes como queda demostrado en [17], por lo que puede ser necesario hacer previamente una estimación del peso o importancia que se debe asignar a cada una de las dimensiones.

3. Una Arquitectura Software para gestionar mediciones de DQ

Volviendo al problema que motivó esta línea de investigación y teniendo en cuenta los contenidos presentados en la sección segunda, en este apartado queremos introducir brevemente una primera aproximación a la solución.

Básicamente nuestra propuesta consiste en una arquitectura consistente en dos capas: una que complementa al modelo de datos y otra que complementa al modelo de procesos del Sistema de Información. Por razones de eficiencia, diseño y seguridad es un complemento no invasivo es decir, en la medida de lo posible, no debe modificar ninguna de ellas.

La capa de datos, a la que llamamos **mapas de agregación**, tiene como objetivo recoger los datos adicionales necesarios para la evaluación. Además de la necesidad de que los datos estén suficientemente auto-descritos, debemos tener en cuenta el requisito tecnológico de la interoperabilidad semántica. Por todo ello pensamos en el uso de tecnologías semánticas ya que permiten un descubrimiento más efectivo, la automatización e integración así como la reutilización de los datos entre distintas aplicaciones [21]. Además de satisfacer los requisitos, aportarían valor añadido a nuestra solución, como la posibilidad de disponer de espacios de anotación coherentes con el espacio de datos.

Por su parte, la capa que complementa al modelo de procesos, a la que llamamos **Capa de Gestión de Lógica de DQ**, debería englobar tanto los procesos técnicos relacionados con la adquisición y mantenimiento de estos datos adicionales, las correspondientes reglas de negocio, y su ligadura a los datos almacenados en las bases de datos organizacionales, como aquellos procesos de gestión orientados a la computación y cálculo de los niveles de calidad de datos de los datos basados en la reglas de negocio. Además también existiría la posibilidad de extrapolar medidas en contextos similares usando razonadores, establecer mecanismos de recomendación de documentos, y la posibilidad de optimizar ciertos procesos típicos del sistema de información atendiendo a criterios de calidad de datos, como puede ser la búsqueda y recuperación de información filtrando aquellos documentos (o conjunto de datos) cuyos niveles de calidad no alcancen los mínimos deseables.

A fin de modelar adecuadamente esta arquitectura software, son varias las tareas que tenemos que acometer: (1) identificar claramente qué datos están afectados por los requisitos de calidad de datos, (2) identificar cuáles son las dimensiones de calidad más adecuadas para los datos y cómo identificar los datos adicionales; (3) definir cómo conseguir valores para los datos adicionales y cómo anotarlos, y finalmente (4) identificar cómo calcular las evaluaciones de calidad de acuerdo a la percepción de DQ de diferentes grupos de agentes usando los modelos de calidad de datos proporcionados.

A continuación presentamos los retos de investigación que nos estamos encontrando a la hora de aplicar esta solución.

En la Web Semántica, también llamada Web de Datos, los datos se modelan como un grafo etiquetado dirigido, donde cada nodo se corresponde con un recurso (sujetos y objetos), y cada arco con un predicado. En una primera aproximación proponemos es hacer anotaciones RDF de los datos adicionales correspondientes a las dimensiones de calidad que son de interés para un determinado grupo de usuarios.

A partir del Modelo de Información de Medición de Calidad de Datos presentado en [4], hemos creado una ontología usando OWL que nos permitirá hacer las anotaciones pertinentes. Una vez acotado el espacio de datos, es preciso acotar el espacio de los elementos anotables según el nivel de granularidad. En [22] los autores identifican los siguientes: RDF Database, Documentos de Web Semántica (SWD) subgrafos RDF o Web Semántica. En esta primera aproximación hemos decidido centrarnos en las sentencias (como si fuesen hechos), por lo que nuestro objetivo será escribir sentencias de calidad de datos sobre sentencias, proceso conocido como reificación [23]. Para identificar las dimensiones de calidad que mejor se adaptan a un contexto, proponemos al lector consultar algunos de los modelos de calidad de datos propuestos en la literatura: por ejemplo para entornos Web se puede consultar el propuesto por Moraga et al en [24].

Por ejemplo, consideremos la sentencia "*Buddy owns a business*" y "*business has-WebSite http://www.budy.com*" cogido directamente de [23] y mostrado en la Figura 2. Imaginemos que alguien quiere conocer por ejemplo cómo de fiable, qué grado de reputación tiene o lo actual que las sentencias son. Para ello se necesitarán datos adicionales como los marcados en la mencionada figura.

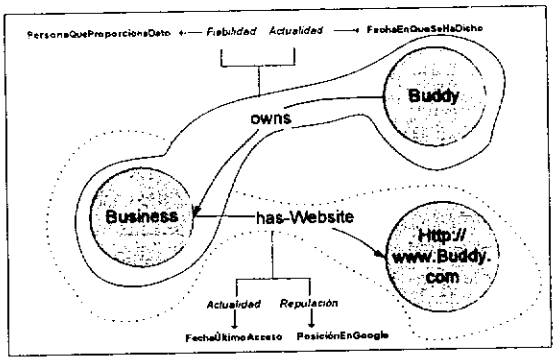

Figura 2. Dimensiones de DQ para diferentes Sentencias, extraído de [24]

Desde un punto de vista práctico/tecnológico aparecen cuatro aspectos claves en la medición de DQ sobre los que estamos trabajando en la actualidad: (1) cómo podemos acceder a los datos almacenados en las bases de datos organizacionales y cómo representarlos respetando tanto el modelo de datos como la legalidad (2) quién debe proporcionar los datos adicionales y cuántos deberían recogerse (3) cómo y dónde deben almacenarse estos valores y (4) cómo se puede conseguir un valor representativo para todos los valores del mismo dato adicional para todos los usuarios a fin de evaluar la calidad.

Según estamos constatando en nuestras investigaciones los principios de DataSpaces [25] y de Linked Data [26], definido como un conjunto de buenas prácticas para la publicación y conexión de datos estructurados en la web, permitirían plantear una respuesta a la pregunta (1). Para el caso de Sistemas de Información con Bases de datos heredadas, disponemos de un sistema de encapsulamiento (*wrapping*) definido en [27], que convierte mediante reingeniería una base de datos heredada en servicios Web, y que al estar basado en transformaciones MDA, y definiendo las adaptaciones oportunas, se podrían expresar de acuerdo a la ontología de medición de DQ creada a tal efecto. Para responder a las preguntas (2) a (4) se están explorando los beneficios de la redes sociales: en [19], los autores introducen formas de almacenar las diferentes opiniones y de operar con ellas para obtener un nivel de calidad.

4. Conclusiones y trabajo futuro

En este artículo hemos descrito brevemente los avances que vamos realizando en nuestra línea de investigación, en la que hemos observado la necesidad de tener en cuenta los aspectos de calidad de datos en la operativa de los sistemas de información para optimizar su rendimiento. Una parte fundamental de este trabajo consiste en poder generar, almacenar y usar datos relativos a la calidad de datos de modo transparente a las aplicaciones, facilitando así la integración y reutilización de dichos datos a través de distintas aplicaciones.

Esta labor, que está siendo desarrollado actualmente, tiene abiertas dos líneas de trabajo: una dedicada a la identificación y aplicación de modelos de calidad de datos a distintos contextos, sobre todo relacionados con la Web (como por ejemplo aplicaciones Web 2.0 y *Mashups*) y por otro el establecimiento y prueba de los mapas de agregación (usando los modelos de calidad obtenidos en la línea paralela), así como la prueba de los procedimientos correspondientes para la adquisición de datos adicionales, y el cálculo de los niveles de calidad.

Agradecimientos. Esta investigación forma parte de los proyectos DQNet (TIN2008-04951-E) y ESFINGE (TIN2006-15175-C05-05) financiada por el MEC, e IQMF-Tool financiada por la UCLM.

Referencias

1. Strong, D.M., Lee, Y.W., Wang, R.Y.: Data Quality in Context. Communications of the ACM **40** (1997) 103-110
2. Wang, R.Y., Madnick, S.: Data Quality Requirements: Analysis and Modelling. Ninth International Conference on Data Engineering (ICDE'93). IEEE Computer Society, Vienna, Austria (1993) 670-677
3. Price, R., Shanks, G.: The Effect of Data Quality Tag Values and Usable Data Quality Tags on Decision-Making. MCIS'09. LNCS, Brisbane, Australia (2009)
4. Caballero, I., Verbo, E.M., Calero, C., Piattini, M.: A Data Quality Measurement Information Model based on ISO/IEC 15939. 12th ICIQ. MIT, Cambridge, MA (2007)

5. Missier, P., Embury, S., Greenwood, M., Preece, A., Jin, B.: Quality views: capturing and exploiting the user perspective on data quality. Proceedings of the 32nd international conference on Very large data bases-Volume 32 (2006) 977-988
6. Stivilia, B.: A Model for Information Quality Change. In: Robbert, M.A., O'Hare, R., Markus, M.L., Klein, B. (eds.): 12th International Conference on Information Quality. MIT, Cambridge, USA (2007) 39-49
7. Lee, Y.W., Pipino, L.L., Funk, J.D., Wang, R.Y.: Journey to Data Quality. Massachussets Institute of Technology, Cambridge, MA, USA (2006)
8. ISO/IEC: ISO/IEC 9126. Software Engineering-Product Quality. Parts 1 to 4. International Organization for Standardization/International Electrotechnical Commission. (2001)
9. Al-Hakim, L.: Procedure for Mapping Information Flow: A case of Surgery Management Process. In: Al-Hakim, L. (ed.): Information Quality Management: Theory and Applications. Idea Group Publishing, Hershey, PA, USA (2007) 168-188
10.Caro, A., Calero, C., Caballero, I., Piattini, M.: A proposal for a set of attributes relevant for Web Portal Data Quality. Software Quality Journal (2008)
11.ISO/IEC-JTC1/SC7: CD 25012.2 Software engineering: Software Quality Requirements and Evaluation (SQuaRE) A" Data Quality Model - N3574- 2006-07-10. International Organization for Standarization (2008)
12.Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques. Springer-Verlag Berlin Heidelberg, Berlin (2006)
13.English, L.: Improving Data Warehouse and Business Information Quality: Methods for reducing costs and increasing Profits. Willey & Sons, New York, NY, USA (1999)
14.Loshin, D.: Enterprises Knowledgement Management: The Data Quality Approach. Morgan Kauffman, San Francisco, CA, USA (2001)
15.Loshin, D.: Data Quality and Business Rules. In: Piattini, M., Calero, C., Genero, M. (eds.): Information and Database Quality. Kluwer Academic Publishers (2001)
16.Wang, R.Y.: A Product Perspective on Total Data Quality Management. Communications of the ACM 41 (1998) 58-65
17.Helfert, M., Foley, O., Ge, M., Cappiello, C.: Limitations of Weighted Sum Measures for Information Quality. AMCIS, San Francisco, CA, USA (2009)
18.Herrera-Viedma, E., Pasi, G., Lopez-Herrera, A.: 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 54 (2006) 538-549
19.Caballero, I., Verbo, E.M., Serrano, M.Á., Calero, C., Piattini, M.: Tailoring Data Quality Models using Social Network Preferences. 2nd International Workshop on Managing Data Quality in Collaborative Information Systems Springer, Brisbane, Australia (2009) 1-15
20.Wang, R., Strong, D.: Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems; Armonk; Spring 1996 12 (1996) 5-33
21.Guha, R., McCool, R., Miller, E.: Semantic search. Proceedings of the 12th international conference on World Wide Web. ACM Press, Budapest, Hungary (2003) 700-709
22.Ding, L., Finin, T., Joshi, A., Peng, Y., Pan, R., Reddivari, P.: Search on the Semantic Web. IEEE Computer 38 (2005) 62-69
23.Daconta, M., Obsrt, L., Smith, K.: The Semantic Web: A guide to the future of XML. Web Services and Knowledge Management. Willey Inc, Indianápolis, Indiana (2003)
24.Moraga, C., Moraga, M., Calero, C., Caro, A.: SQuaRE-Aligned Data Quality Model for Web Portals. QSIC'09. IEEE, Jeju (2009)
25.Halevy, A., Frankling, M., Maier, D.: Principles of DataSpaces Systems. PODS (2006)
26.Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. Special Issue on Linked Data, International Journal on Semantic Web and Information System (2009) To appear
27.Pérez-Castillo, R., García-Rodríguez , I., Caballero, I.: PRECISO: a reengineering process and a tool for database modernisation through web services. SAC'09 (2009) 2126-2133