Reflexiones sobre la calidad de la información y el diseño de bases de datos[1]

Ismael Caballero, Mario Piattini, Marcela Genero, Coral Calero, Francisco Ruiz

Ponente: Mario Piattini

Grupo ALARCOS

Escuela Superior de Informática. Universidad de Castilla-La Mancha.

Ronda de Calatrava 7. 13071, Ciudad Real. España.

E-mail: icaballero@ole.com, mpiattin@inf-cr.uclm.es.

Tel. +34 926 29 53 00 (ext. 3715). Fax + 34 926 29 53 54.

Resumen

Para ser más productivas las empresas deben tratar la información como un activo más. Esto hace necesario controlar la calidad de los datos organizacionales para un mejor aprovechamiento de los sistemas de información implantados. La calidad es un concepto multidimensional. Distintos autores dedicados al tema, tratan de analizar las distintas dimensiones de la calidad desde puntos de vista tales como las necesidades de los usuarios o la modelización de la realidad para diseñar el Sistema de Información. La idea es ampliar las metodologías tradicionales para que recoja las necesidades de los usuarios y sus requisitos de calidad de datos representados por ciertas dimensiones y los incorpore al diseño conceptual y a partir de ahí seguir con la metodología tradicional. Una vez que tenemos implementada nuestra base de datos con las dimensiones especificadas, podremos implementar mecanismos en un Lenguaje de Manipulación de Datos para la calidad de los datos que estamos usando. Pero deben ser las empresas las que tomen la decisión de planificar y controlar la calidad de sus datos. Este documento presenta una serie de razones por las que las empresas deben acometer la calidad de los datos, un análisis de las dimensiones de calidad según varios autores y una metodología para el diseño de bases donde se incorpore la calidad de los datos.

1. Introducción

En la actualidad la mayor parte de las empresas se enfrentan a un grave problema de "polución de datos", en efecto, disponen de demasiados datos, debido principalmente a tres motivos:

- La facilidad y el bajo coste con el que se pueden capturar datos gracias a la mejora y difusión de las tecnologías de entrada de datos: códigos de barra, OCR (reconocedores ópticos de caracteres), tarjetas de cliente, tarjetas de crédito, ... A lo que habría que añadir la gran cantidad de datos que se puede obtener por medio de Internet.

- La redundancia incontrolada de datos. Como sabemos, a nivel físico la redundancia puede ser razonable y necesaria por motivos de eficiencia, pero a nivel lógico deberían implementarse los mecanismos necesarios para mantener sincronizados los datos redundantes. Desafortunadamente, y debido al propio funcionamiento diario, los sistemas de información crecen de manera desordenada y poco planificada, no existiendo en muchos casos una arquitectura de información en la empresa.

- La existencia de grandes cantidades de datos históricos "caducados", que ya no sirven para realizar ningún proceso ni obtener ningún tipo de información relevante. Como señala Orr (1998), al igual que sucede con los miembros de un organismo biológico, los datos que no se utilizan terminan atrofiándose.

Esta polución puede llegar a tener graves consecuencias; así, por ejemplo, Celko (1995) afirma que la mitad del coste total de implementar un almacén de datos (*datawarehouse*) puede deberse a una pobre calidad de datos. El Gartner Group ha advertido también que la pobre calidad de datos ha sido una de las causas de fracaso más importantes en los proyectos de reingeniería. Se hace por tanto imprescindible para el buen funcionamiento del Sistema de Información de la empresa abordar el tema de la calidad de los datos, para que éstos se conviertan en verdadera información y conocimiento. Las empresas deben gestionar la información como un producto importante, capitalizar el conocimiento como un activo principal y, de esta manera, sobrevivir y prosperar en la economía digital (Huang et al., 1999). Mejorando la calidad de la información se conseguirá mejorar la satisfacción de los clientes y, al mismo tiempo, la satisfacción del personal, lo que hará mejorar la empresa en su conjunto.

Desafortunadamente, hasta hace muy poco tiempo, los aspectos de la calidad se han centrado en la calidad de los programas, descuidándose el aspecto de la calidad de los datos (Sneed y Foshag, 1998). Incluso en el diseño tradicional de las bases de datos, los aspectos referidos a la calidad no se han incorporados explícitamente (Wang et al., 1993). Sí es cierto que aunque la investigación y la práctica en bases de datos tradicionalmente no han estado centradas en temas relativos a la calidad, muchas de las herramientas y técnicas desarrolladas (restricciones de integridad y teoría de la normalización, gestión de transacciones, etc.) han tenido influencia en la calidad. Nosotros opinamos que ha llegado el momento de considerar la calidad de la información como un objetivo principal a perseguir, más que como hasta ahora como un subproducto del proceso de creación y desarrollo de bases de datos.

Existen, en general, dos aspectos a tener en cuenta en la calidad de la información: la calidad de la base de datos en su conjunto y la calidad de la presentación de los datos. En efecto, es muy importante que los datos de la base de datos reflejen correctamente el mundo real, esto es, que sean precisos; pero también que se puedan entender fácilmente. Por lo que se refiere a la calidad de la base de datos en su conjunto, depende de tres "calidades": la del SGBD (Sistema de Gestión de Bases de Datos) utilizado, la del modelo[2] de datos (tanto conceptual como lógico) y la de los propios datos.

En esta comunicación nos centraremos en las características relevantes de los propios datos, por lo que respecta a la calidad de los modelos de datos, remitimos al lector interesado a Piattini et al. (1999).

2. Dimensiones de la calidad de la información

Como sabemos, el concepto de calidad es relativo, está en los ojos del observador, por lo que podemos considerar la calidad como un concepto multidimensional, sujeta a restricciones y ligada a compromisos aceptables (Piattini et al., 1997). Varios autores han

propuesto en estos últimos años diferentes dimensiones para la calidad de los datos; así, por ejemplo, Redman (1996) considera tres tipos de categorías en las que se pueden agrupar las dimensiones de la calidad (que son autoexplicativas):

- Dimensiones de Calidad de la Vistas de Datos.
- Contenido: Relevancia de los datos, Obtenibilidad de los valores y claridad de definición.
- Alcance: Comprensividad y Esencialidad.
- Nivel de Detalle: Granularidad de los atributos y Amplitud del Dominio.
- Composición: Naturalidad, Identificalbilidad, Homogenedidad, Redundancia Innecesaria Mínima.
- Consistencia de las Vistas: Consistencias Estructurales y Semánticas.
- Reacción al Cambio: Flexibilidad y Robustez.
- Dimensiones de Calidad de los Valores de Datos.
- Exactitud
- Completitud
- Actualidad
- Consistencia de Valores
- Dimensiones de Calidad de la Representación de los datos.
- Ser apropiados
- Interpretabilidad
- Portabilidad
- Precisión del Formato
- Flexibilidad del Formato
- Capacidad de representar valores Nulos.
- Uso Eficiente del espacio de almacenamiento
- Consistencia de Representación

English (1999), por su parte, destaca dos tipos de cuestiones relativas a la calidad de los propios datos:

- Calidad inherente, es decir la precisión de los datos, el grado en que los datos reflejan

exactamente los objetos del mundo real que representan, que abarcaría: conformidad con la definición, compleción de valores, validez o conformidad con las reglas del negocio, precisión respecto a la fuente, precisión respecto a la realidad, no duplicación, accesibilidad

- Calidad pragmática, el grado en que los datos permiten a los "trabajadores del conocimiento" satisfacer los objetivos de la empresa de forma eficaz y eficiente: oportunidad, claridad contextual, integridad de derivación, usabilidad, corrección o compleción de hechos.

En Wand y Wang (1996) se analizan, desde principios ontológicos, algunas de las causas de la mala calidad de los datos debida a deficiencias en el diseño, identificando cuatro dimensiones de calidad, véase tabla 1.

Dimensión Calidad de Datos	**Naturaleza de la deficiencia**	**Fuente de la deficiencia**
Compleción	Representación impropia: Estados del SI[3] ausentes	Fallo en el diseño
No ambigüedad	Representación impropia: Varios estados del MR[4] mapeados al mismo estado SI	Fallo en el diseño
Significación	Estados SI sin sentido y confusión: mapeo a estado sin sentido	Fallo en el diseño y fallo en la operación
Corrección	Confusión: mapeo a un estado incorrecto	Fallo en la operación

Tabla 1.- Calidad de los datos y deficiencias de diseño

Como señalan estos autores, se pretende que cada estado del mundo real, corresponda de forma unívoca con un estado del Sistema, En caso de que no se verifique la relación unívoca o bien que al operar con los datos no se obtengan los valores esperados se produce una "*deficiencia*" de datos. Estas deficiencias pueden ser de distinta naturaleza:

- **Deficiencias de Diseño: Se produce una anomalía de este tipo cuando hay estados del mundo real que no se corresponden con un único estado correcto del sistema de Información o viceversa. Nos encontramos con los siguientes tipos:**
  - **Representación incompleta: cuando hay estados del mundo real, que perteneciendo a la semántica de nuestro problema, se quedan sin representación en el SI.**
  - **Representación ambigua: se produce cuando dos o más estados del mundo real son representados por el mismo estado del SI.**
  - **Estados sin sentido: ocurre cuando aparecen en el SI estados, que no están asociados a ningún estado del mundo real.**
- **Deficiencias de Operación: Este tipo de deficiencias se da cuando el funcionamiento del sistema sobre una colección de datos, no produce la salida esperada; por ejemplo, cuando**

hay operaciones en el SI que partiendo de un estado dan como resultado el estado de la realidad que le correspondería, o bien, que partiendo de un estado de la realidad, llevan a un estado del SI que no es el que le representa.

- **Deficiencias Relacionadas con la descomposición: Muchas veces para diseñar un SI se recurre a la estrategia *"Divide y Vencerás"* para hacer más sencillo el problema. En algunos casos esta "división" se hace mal y provoca la aparición de cualquiera de las deficiencias de diseño vistas anteriormente.**

Por último, cabe destacar las cuatro categorías de calidad de datos definidas por Strong et al. (1997a), para cada categoría estos autores identifican varias dimensiones:

- **Intrínseca: precisión, objetividad, credibilidad, reputación**
- **Accesibilidad: accesibilidad, seguridad de acceso**
- **Contextual: relevancia, valor añadido, oportunidad, compleción, cantidad de datos**
- **Representacional: interpretabilidad, facilidad de comprensión, representación concisa, representación consistente**

3. Diseño de bases de datos y calidad de datos

Orman et al. (1994) sugiere tres enfoques para mejorar la calidad intrínseca de la base de datos:

- Construir modelos semánticos más ricos que reflejen mejor la realidad.
- Reforzar las bases de datos con un mayor número de restricciones para identificar y discriminar datos con problemas y enlazar dichos datos con las aplicaciones apropiadas.
- Restringir el uso de los datos a procesos predefinidos, no permitiendo que sean modificados por cualquier proceso para que no puedan ser borrados accidentalmente.

Aunque todos estos enfoques permiten mejorar la calidad de los datos, no son suficientes ya que es necesario un soporte adecuado para gestionar de alguna forma las dimensiones de la calidad (Wang et al., 1992). Desafortunadamente existen pocas propuestas que contemplen la calidad de los datos como factor fundamental en el proceso de diseño. Los trabajos de Wang et al. (1993) y (1995) constituyen una excepción en este sentido. Estos autores proponen una metodología que complementa a las tradicionales metodologías de diseño de bases de datos (De Miguel et al., 1999), véase figura 1.

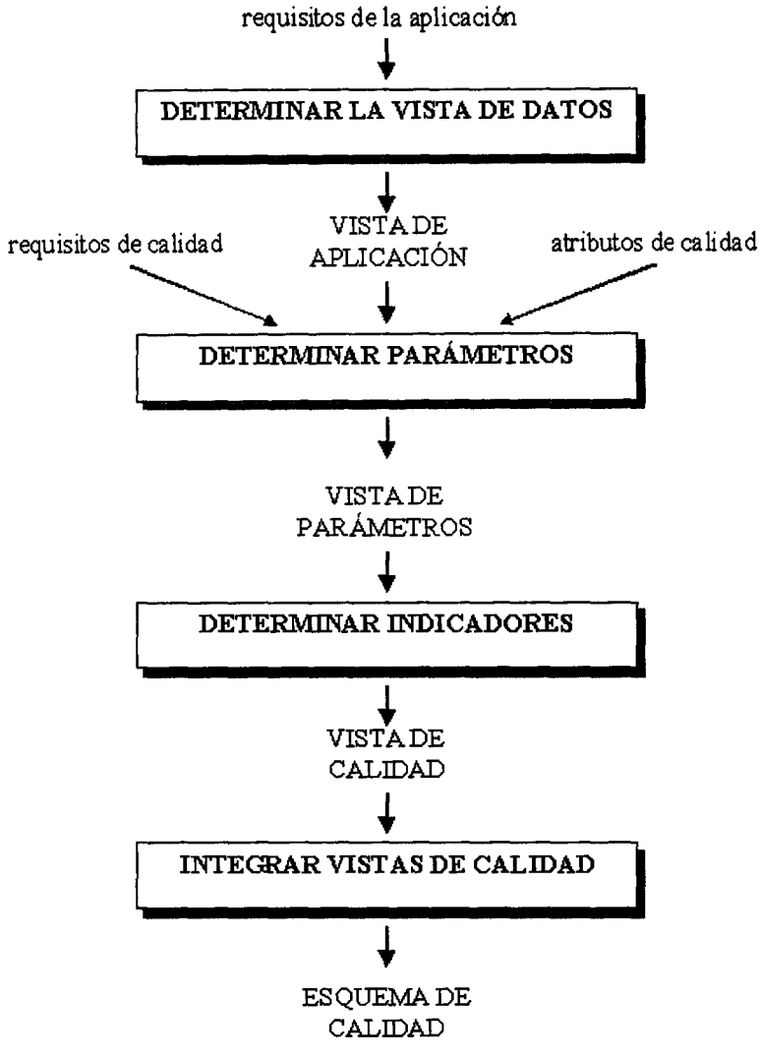

Figura 1.- Calidad en el diseño de bases de datos (Wang et al., 1993)

En la primera etapa, véase figura 1, además de crear el esquema conceptual, por ejemplo, utilizando el modelo entidad/interrelación, se deberían identificar los requisitos de calidad y los atributos candidatos; determinando, a continuación, la "vista de parámetros de calidad" así, a cada elemento del esquema conceptual se le puede asociar un parámetro de calidad. Por ejemplo, en una base de datos académica, al atributo "nota de selectividad" se podría asociar la precisión y la oportunidad. Posteriormente se objetiva los parámetros subjetivos añadiendo etiquetas a los atributos del esquema conceptual (la fuente, para conocer la precisión, y la fecha, para la oportunidad, de las notas de selectividad) y se integran las diferentes vistas.

Por otro lado, también se propone extender las bases de datos relacionales con indicadores que permitan asignar estos parámetros objetivos y subjetivos a la calidad de los valores de la base de datos (Wang et al., 1995). Así, por ejemplo, en la tabla que se muestra a continuación se almacena para cada valor de la base de datos la fuente que proporciona el dato y la fecha en la que lo hace, además se debería conocer la credibilidad de la fuente (que en este caso podría ser alta), lo que ayudaría a los "trabajadores del conocimiento" en la toma de decisiones.

ALUMNO	NOTA SELECTIVIDAD	NOTA MEDIA CARRERA
William Smith	8 <30/10/90, MEC>	7 <30/7/95, Escuela Inf.>
Gene Hackman	9 <30/10/90, MEC>	6 <10/9/96, Escuela Inf.>
	. . .	. . .

4.- Conclusiones y trabajos futuros

Podemos afirmar que al igual que en estos últimos años la calidad de los productos y servicios ha sido un factor esencial para el éxito de las empresas, en la próxima década lo será la calidad de la información.

Si realmente consideramos que la información constituye el activo más importante de las organizaciones, una de las primeras funciones de los profesionales de las TI debería ser asegurar la calidad de la misma. Hemos presentado algunas propuestas reciente para caracterizar y asegurar la calidad de la información, pero es necesario profundizar también en la investigación sobre la calidad de los procesos asociados a la información: el proceso de modelado, el proceso de obtención y carga de los datos, y el proceso de presentación.

Las organizaciones tendrían que, por un lado, definir una política de calidad (véase, por ejemplo, Redman, 1996) que establezca las obligaciones de cada función con el fin de asegurar la calidad de los datos en todas sus dimensiones; mientras que por otro deberán implementar un proceso con el fin de evaluar la calidad de la información de que disponen. Existen varias propuestas para evaluar la calidad de información, entre las que destacamos la TQdM (*Total Quality data Management*) de English (1999).

Un aspecto crítico para poder llevar a cabo este proceso de evaluación es la definición de unas métricas que sean significativas, y que permitan realmente analizar y mejorar la calidad. En Huang et al. (1999) se proponen tres tipos de

métricas: subjetivas (basadas en el juicio de los usuarios de los datos), objetivas independientes de la aplicación (como la corrección) y objetivas dependientes de la aplicación (específicas para un dominio determinado). Además, se debería medir el valor de la información, tanto el producido por sistemas operacionales como los de ayuda a la toma de decisiones. La forma de medir el valor de la información en estos dos tipos de sistemas varía considerablemente.

Referencias

Celko, J. (1995). Don`t Warehouse Dirty Data. *Datamation* , 15 octubre, 42-52.

De Miguel, A., Piattini, M. y Marcos, E. (1999). Diseño de bases de datos relacionales. Madrid, Ra-Ma.

English, L. (1999). Improving Data Warehouse and Business Information Quality. John Wiley & Sons, Inc.

Huang, K-T., Lee, Y.W. y Wang, R.Y. (1999) *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River.

Orman, L., Storey, V. Y Wang, R. (1994). Systems Approaches to Improving Data Quality. TDQM-94-05, Agosto 1994. Disponible en http://www.

Orr, K. (1998). Data Quality and System Theory. Communications of the ACM, 41 (2), 66-71.

Piattini, M., Calvo-Manzano, J.A., Cervera, J. y Fernández, L. (1997). *Análisis y diseño detallado de aplicaciones informáticas de gestión*. Madrid, Ra-Ma.

Piattini, M., Genero, M y Calero, C. (1999). Calidad de esquemas conceptuales. *Novática*, julio/agosto 1999.

Redman, T. C. (1996). *Data Quality for the Information Age* . Artech House, Boston.

Sneed, H.M. y Foshag, O. (1998): Measuring Legacy Database Structures. *Proc. of The European Software Measurement Conference FESMA'98* , Coombes, Hooft and Peeters (eds.), 199-210.

Storey, V. C. Y Wang, R. (1994). Modeling Quality Requirements in Conceptual Database Design. TDQM-94-02, Mayo 1994.

Strong, D.M., Lee, Y.W. y Wang, R.Y. (1997a). Data Quality in Context. Communications of the ACM 40 (5), 103-110.

Strong, D.M., Lee, Y.W. y Wang, R.Y. (1997b). 10 Potholes in the Road to Information Quality. *IEEE Computer* 38-46.

Wand, Y. y Wang, R.Y. (1996). Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, 39 (11), 86-95.

Wang, R. Y., Kon, H. B. y Madnick, S. E. (1993). Data Quality Requirements Analysis and Modeling. *Proc. of the 9th International Conference on Data Engineering* , Viena, IEEE Computer Society, 670-677.

Wang, R.Y., Reddy, M.P. y Kon, H.B. (1995). Toward quality data: An attribute-based approach. *Decision Support Systems*, 13, 349-372.

Notas

[1] Este trabajo se encuadra dentro del proyecto CALIDAT, que desarrolla la empresa Cronos Ibérica, S.A. en colaboración con la Universidad de Castilla-La Mancha con financiación de la Consejería de Educación y Cultura de la Comunidad de Madrid (Ref: 09/0013/1999).

[2] Entendido como esquema, no como facilidad o técnica de modelado.

[3] SI = Sistema de Información

[4] MR = Mundo Real