

Medición, evaluación y certificación de calidad de datos en la transformación digital

F. Gualo ^(1,2), I. Caballero ⁽²⁾, M. Piattini ⁽²⁾, J. Verdugo ⁽³⁾ M. Rodríguez ^(2,3)

⁽¹⁾ DQTeam: ITSI, Camino de Moledores SN, Ciudad Real, 13005.

⁽²⁾ Grupo Alarcos, Universidad de Castilla-La Mancha, Paseo de la Universidad 4, 13005.

⁽³⁾ AQCLab: ITSI, Camino de Moledores SN, Ciudad Real, 13005.

Teléfono.: 659183745 Correo electrónico: fgualo@dqteam.es

RESUMEN: Los datos están en el centro del negocio de muchas organizaciones como uno de los activos más importantes, ya que las decisiones están basadas en datos. Por ello, las organizaciones necesitan confiar en sus datos. Una actividad que ayuda a conseguir la fiabilidad de los datos es la evaluación y certificación del nivel de calidad de los repositorios de datos de las organizaciones. Este artículo describe los resultados de la aplicación de un proceso de evaluación y certificación de la calidad de los datos a los repositorios de tres organizaciones europeas pertenecientes a diferentes sectores. Presentamos los resultados desde el punto de vista tanto del equipo de evaluación de la calidad de los datos como de las organizaciones que se sometieron al proceso de evaluación. En este sentido, las organizaciones implicadas han reconocido explícitamente varios beneficios tras conseguir la certificación de calidad de datos para sus repositorios (por ejemplo, sostenibilidad organizativa a largo plazo, mejor conocimiento interno de los datos y una gestión más eficiente de la calidad de los datos).

1. INTRODUCCIÓN

La mayoría de las organizaciones hacen uso de ellos para sus actividades operativas, tácticas, y estratégicas independientemente del sector y la zona geográfica. En consecuencia, junto con las personas, los datos pueden considerarse como uno de los activos más importantes para las organizaciones [1], [2]. Esta percepción de los datos como un activo implica que las organizaciones deben comprometerse firmemente con la idea de la calidad de los datos, ya que cuanto mejores sean sus datos, mayores serán los beneficios que puedan obtener de su uso [3], [4]. Por lo tanto, se puede afirmar que los datos con niveles adecuados de calidad pueden permitir nuevas formas de innovar en los negocios en un mercado cada vez más competitivo [5]. Distintos expertos como Olson [6], Redman [7], o Laney [8] demostraron la conexión entre la mala calidad de los datos – más de 3.100 millones de dólares en pérdidas en EEUU – y el aumento del coste y la complejidad en el desarrollo de sistemas [8], afirmando que:

- La mala calidad de los datos es una de las principales razones por las que el 40% de las iniciativas empresariales no alcanzan su objetivo de beneficios.
- La calidad de los datos afecta a la productividad laboral global hasta en un 20%.
- A medida que se automatizan más procesos empresariales, la calidad de los datos se muestra como el factor limitante de la calidad global de los procesos.

Teniendo en cuenta todo lo anterior, es evidente que las organizaciones deben invertir los recursos adecuados en el despliegue de mecanismos que ayuden a garantizar la fiabilidad de los datos, es decir, que éstos tengan el nivel de calidad adecuado para su uso [9]. El establecimiento de estos mecanismos para asegurar y controlar la calidad de los datos es una

tarea que debe planificarse con suficiente antelación, y debe llevarse a cabo con objetivos claros y de conformidad con la estrategia de la organización [10], [11]. A la luz de nuestra experiencia en el campo de la certificación industrial de la calidad del software [12], proponemos que, de forma análoga, la certificación del nivel de calidad de los repositorios de datos específicos puede proporcionar a las organizaciones la confianza necesaria en esos datos. En [13], [14] se presenta el entorno establecido para la evaluación de la calidad de los datos. Este entorno consta de dos elementos principales:

- Un modelo de calidad de datos basado en ISO/IEC 25012 [15] e ISO/IEC 25024 [16], con las características y propiedades de calidad de los datos.
- Un proceso de evaluación de calidad de los datos basado en ISO/IEC 25040 [17], que define las actividades, tareas, etc. para la evaluación.

Este entorno de evaluación de la calidad de los datos ha sido desarrollado por AQCLab, el primer laboratorio acreditado por ENAC /ILAC para la evaluación de la calidad de los productos de software y de los datos basada en las normas ISO/IEC 25000. Los informes de evaluación resultantes de las evaluaciones de calidad de datos realizadas a los repositorios de datos con este entorno son el principal insumo para el proceso de certificación de calidad realizado por AENOR (entidad certificadora de TIC líder en España y Latinoamérica). Este artículo presenta una visión general del entorno, y diferentes experiencias industriales sobre la aplicación del entorno en la evaluación de la calidad de los datos en organizaciones, así como los hallazgos que descubrimos a partir de estas experiencias.

2. DESARROLLO/DESCRIPCIÓN

El **modelo de calidad de datos** está compuesto por el conjunto de características y propiedades de calidad. Este, está basado en ISO/IEC 25012 [15] e ISO/IEC 25024 [16], de manera que dispone de características que aportan distintos puntos de vista en relación a la calidad de los datos, y propiedades, que aportan un mayor detalle de cada una de las características (véase Fig. 1). Este modelo está basado en las características inherentes de ISO/IEC 25012, es decir las características que cumplen todos los datos por su mera existencia: exactitud, completitud, consistencia, credibilidad y actualidad.



Fig. 1. Visión general del modelo de calidad de datos.

Para cada una de las características es posible definir un valor de calidad en el rango [1,5] a partir de la medición y agregación de las propiedades de una característica. El valor de la medición de una propiedad de calidad se obtiene mediante la aplicación de un determinado

método de medición. La medición de cada propiedad se normaliza dentro del intervalo [0,100], y se le asigna un determinado nivel siguiendo los intervalos mostrados en la Tabla 1.

Valor de medición de la propiedad	Nivel	Descripción
0-25	1	Calidad deficiente
25-50	2	Calidad insuficiente
50-75	3	Buena Calidad
75-95	4	Muy buena calidad
95-100	5	Calidad excelente

Tabla 1. Niveles de calidad y descripción para las características de calidad

Para asignar un valor de calidad a cada característica, es necesario realizar una agregación del valor de la medición de cada una de las propiedades de calidad que se consideran para la característica a la que pertenecen. Esta agregación se realiza mediante una función por perfiles que define un conjunto de rangos, que especifican el número máximo de propiedades de calidad de los datos admisibles en cada nivel de calidad para determinar si la característica de calidad se encuentra en ese rango. A continuación, el perfil del sistema para una característica calculada considerando la cantidad de propiedades de calidad en cada nivel de calidad, se compara con esos rangos, lo que permite determinar el nivel de calidad alcanzado para la característica de calidad.

Por último, se presenta el modelo de certificación de calidad de datos. Este es similar al presentado en [12], y consta de los siguientes pasos:



Fig. 2. Visión general del proceso de certificación de calidad de datos.

En primer lugar, la organización interesada en certificar los niveles de calidad de los datos de un repositorio se pone en contacto con un laboratorio acreditado (como AQCLab) y, tras establecer el alcance de la evaluación de la calidad de los datos, firman un contrato en el que se establecen los términos de la colaboración. Una vez firmado el contrato por ambas partes, comienza el proceso de evaluación. Una vez finalizada la evaluación, el laboratorio acreditado emite el informe de evaluación en el que se detallan los resultados relativos a la calidad de los datos de las características evaluadas. En función de estos resultados, la organización debe decidir si quiere mejorar la calidad del repositorio de datos o iniciar el proceso de certificación si ha alcanzado los valores de nivel de calidad requeridos y deseados. Cuando la organización considera que está preparada para certificar su repositorio de datos, debe contactar con la entidad certificadora para solicitar la certificación de la calidad de los datos, facilitándole la referencia de su informe de evaluación.

A continuación, el organismo de certificación se pone en contacto con el laboratorio acreditado que ha realizado la evaluación para verificar los resultados del informe referenciado por la organización. El laboratorio acreditado proporciona al organismo de certificación la información solicitada. Tras comprobar la validez de los resultados de la evaluación, la entidad de certificación realiza una auditoría de confirmación sobre la organización y su repositorio, y finalmente emite y otorga a la organización el correspondiente certificado de calidad de datos. Este último paso realizado por la entidad certificadora tiene un coste preestablecido y una duración de dos días, independientemente de las características del repositorio de datos a certificar.

3. RESULTADOS Y DISCUSIÓN

Esta sección presenta los resultados de la evaluación de calidad de datos para tres casos de la industria. En este caso hay 3 organizaciones: administración pública, viajes y educación. El objetivo de todas ellas era la evaluación y certificación de todas las características de calidad de datos. Para ello se llevaron a cabo dos evaluaciones, una inicial en la que se identificaron los distintos problemas relacionados con la calidad de los datos, y una posterior a la etapa de mejora de calidad de datos, en la cual, la organización en cuestión resuelve los distintos problemas identificados y es capaz de afrontar con éxito la certificación. El resumen de la información para cada una de las evaluaciones de las organizaciones se puede ver en la Tabla 2.

		Organización 1	Organización 2	Organización 3
Base de datos		Oracle 18c	Teradata 15.0	SQLServer 2017
Número de entidades de datos		46	30	24
Volumen de datos		750 millones	200 millones	100 millones
Características evaluadas		<i>Exactitud, Completitud, Consistencia, Credibilidad, Actualidad</i>		
Reglas por característica	<i>Exactitud</i>	189	94	89
	<i>Completitud</i>	131	100	78
	<i>Consistencia</i>	340	176	91
	<i>Credibilidad</i>	72	48	54
	<i>Actualidad</i>	81	70	63
Total de reglas de negocio		813	488	375

Tabla 2. Resumen de las experiencias

Se aplicó el proceso de evaluación obteniendo los siguientes resultados para cada una de las características en las tres evaluaciones (véase Fig. 3).

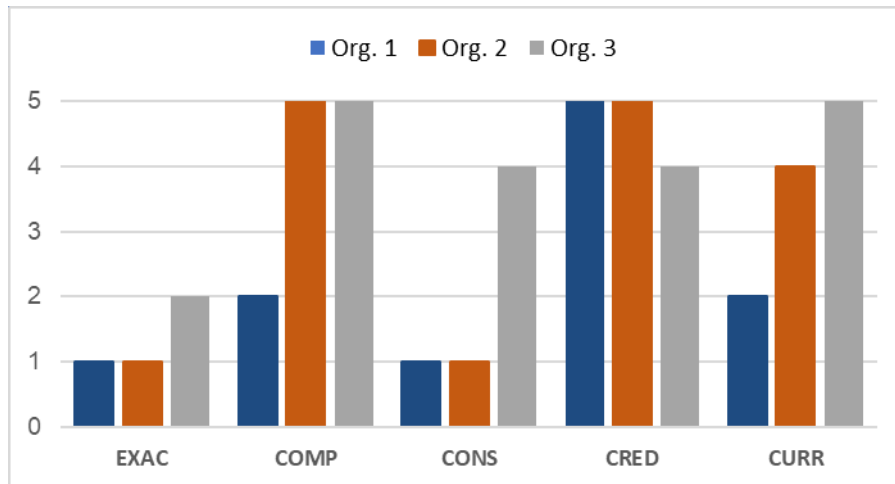


Fig. 3. Resultados de evaluación de calidad de datos para las tres organizaciones

A partir de los resultados de la evaluación de calidad de datos en las tres organizaciones, se identificaron un conjunto de fortalezas y debilidades comunes a todas las evaluaciones. Estas se han resumido y clasificado por propiedad en la Tabla 3.

Característica	Fortalezas	Debilidades
Exactitud	-	Los valores de los datos deben cumplir con ciertas expresiones regulares definidas. Los valores de los datos deben estar definidos dentro de los rangos de aceptabilidad del dominio de valores definido para los atributos.
Compleitud	Los archivos de datos contienen un volumen adecuado de registros para representar las entidades requeridas. Los atributos de los datos contienen los valores necesarios para las diferentes entidades representadas por los datos.	Es necesario revisar las definiciones de datos completos, puesto que hay muchos casos en que no se dispone de valores para todos los atributos que son necesarios.
Consistencia	Las referencias a valores se hacen de manera correcta	Hay casos en los que no se cumple con determinados formatos para datos de similar naturaleza o tipo
Credibilidad	La credibilidad de los datos suele ser alta debido a que se llevan a cabo procesos de integración de datos	Es necesario revisar el nivel de calidad de las fuentes que nutren los repositorios de la organización, debido a que al estar fuera la calidad puede no ser la correcta
Actualidad	Los datos por norma general se actualizan en los tiempos debidos para disponer siempre de una versión actual	Es necesario revisar eventos de actualización de datos, así como los diferentes caminos.

Tabla 3. Niveles de calidad y descripción para las características de calidad

Cada organización en función de sus recursos planificó distintas acciones de mejora para paliar las debilidades encontradas durante el proceso de evaluación, y una vez realizadas las modificaciones pertinentes, se llevó a cabo una nueva evaluación que concluyó con la certificación de los tres repositorios de datos. Durante el proceso de evaluación de calidad de datos surgieron de manera sistemática, una serie de preocupaciones y conclusiones que fueron informadas al equipo de evaluación, y que a continuación se resumen:

- El personal de las organizaciones evaluadas no suele estar suficientemente familiarizado con el concepto de "requisitos de datos", y dado que es algo fundamental al ser la base de la evaluación, se proporciona una plantilla con algunos ejemplos que ayuda a su identificación de manera sistemática.
- Muchas organizaciones a lo largo de mantenimientos en sus sistemas de información habilitaban mecanismos para el cumplimiento de requisitos de datos de manera inconsciente. El punto positivo, es que toda esta parte quedaba documentada y trazada en el tiempo.
- Los resultados de la evaluación hacen ver como las organizaciones son más conscientes de la integridad de datos y actualidad de datos, por lo tanto, los resultados de la evaluación para estas características y sus propiedades suelen ser más altos que para las demás.
- Algunos sistemas de información de las organizaciones no son lo suficientemente potentes para soportar el estrés de consultas en periodos cortos de tiempo, y ha sido necesario virtualizar o sacar de sistemas productivos algunas partes de repositorios.
- Tras la ejecución de distintas evaluaciones, se determina que varias organizaciones tienen problemas de calidad de datos similares tales como: valores de los datos que no cumplen con las reglas de sintaxis específicas, o existencia de registros cuyos valores de datos de algunos atributos no pertenecen a los rangos esperados o permitidos.
- Las organizaciones tienden a definir valores por defecto no eficientes en sus repositorios de manera que en muchos casos dificultan el cruce entre entidades de sus repositorios de datos.
- Los jefes de TI e informática de las organizaciones evaluadas reportan que la identificación de problemas de calidad de datos les ha sido útil para definir planes de mitigación y extrapolar controles a todo el sistema de información, no solo al repositorio. Del mismo modo, han reconocido que, pese al esfuerzo, disponer en un único site del conjunto de requisitos de su repositorio, les aporta valor de cara a su mantenimiento.
- A partir de los resultados y la identificación de problemas de calidad de datos, ha sido posible definir mecanismos de control y aseguramiento de calidad que habilita a que la calidad de los datos nuevos integrados en sus sistemas de información se mantenga, y no se reduzca.

4. CONCLUSIONES

Hay que reconocer que el potencial que la analítica aporta a las organizaciones en esta era de la transformación digital, y la consiguiente necesidad de datos, ha provocado un crecimiento explosivo del interés de las organizaciones por adquirir y procesar grandes cantidades de datos. Siendo los datos el núcleo de las organizaciones actuales, han sido elevados a la categoría de "activo organizativo". Y, al igual que ocurre con otros activos importantes, las organizaciones deben tomar conciencia de la importancia de su calidad a lo largo del tiempo. Esta conciencia implica la necesidad de evaluar sistemáticamente la calidad de los datos. Para apoyar mejor el proceso de evaluación de la calidad de los datos, se necesitan marcos de trabajo. En este artículo, hemos esbozado un entorno de evaluación de la calidad de los datos basado en las normas internacionales ISO/IEC 25012 (que define las características de la calidad de los datos), ISO/IEC 25024 (que define las propiedades de la calidad de los datos y la correspondiente información necesaria para adaptar sus medidas subyacentes), e ISO/IEC 25040 (que define la estructura y los fundamentos del proceso de evaluación que deben adaptar los evaluadores).

Tras aplicar el proceso de evaluación a los repositorios de datos en distintos casos de uso, hemos analizado nuestra experiencia como equipo evaluador y las conclusiones comunicadas por las organizaciones evaluadas. Cabe destacar los tres siguientes beneficios reconocidos por las organizaciones implicadas:

- La evaluación y certificación de la calidad de los datos ayuda a garantizar la sostenibilidad de la organización a largo plazo.
- La evaluación y certificación de la calidad de los datos ayuda a conocer mejor los aspectos internos (de gestión y tecnológicos) de la empresa y las formas de trabajo de la organización.
- El conocimiento adquirido a través del proceso de evaluación y certificación de la calidad de los datos ayuda a respaldar mejor las futuras iniciativas de gestión de la calidad de los datos de la organización.

A partir de las conclusiones de las experiencias, hemos derivado algunas buenas prácticas destinadas a mejorar la eficiencia del proceso de evaluación de la calidad de los datos, así como la satisfacción futura de otras empresas que puedan someterse a una evaluación de la calidad de los datos, facilitando el proceso y haciendo que los resultados obtenidos en el proceso de evaluación de la calidad de los datos sean más útiles para ellas.

Otra conclusión a la que hemos llegado es que la evaluación continua de la calidad de sus datos lleva a las organizaciones a un conocimiento más profundo de sus procesos de negocio, y esto, a su vez, conduce a un mejor rendimiento de la organización.

5. REFERENCIAS

Aquí se catalogarán las referencias a artículos, libros u otros documentos, en la forma:

- [1] R. Y. Wang, «A Product Perspective on Total Data Quality Management», *Commun ACM*, vol. 41, n.º 2, pp. 58-65, feb. 1998, doi: 10.1145/269012.269022.
- [2] P. Woodall, A. K. Parlikad, y L. Lebrun, «Approaches to Information Quality Management: State of the Practice of UK Asset-Intensive Organisations», en *Asset Condition, Information Systems and Decision Models*, Springer, London, 2012, pp. 1-18. doi: 10.1007/978-1-4471-2924-0_1.
- [3] M. Fleckenstein y L. Fellows, *Modern Data Strategy*. Springer International Publishing, 2018. Accedido: 3 de mayo de 2019. [En línea]. Disponible en: <https://www.springer.com/us/book/9783319689920>

- [4] R. Mahanti, *Data Quality: Dimensions, Measurement, Strategy, Management, and Governance*. ASQ Quality Press, 2019.
- [5] R. Sherman, *Business Intelligence Guidebook: From Data Integration to Analytics*. Newnes, 2014.
- [6] J. E. Olson, *Data quality: the accuracy dimension*. Elsevier, 2003.
- [7] T. C. Redman, «Bad Data Costs the U.S. \$3 Trillion Per Year», *Harvard Business Review*, 22 de septiembre de 2016. Accedido: 30 de noviembre de 2020. [En línea]. Disponible en: <https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year>
- [8] D. B. Laney, *Infonomics: how to monetize, manage, and measure information as an asset for competitive advantage*. Routledge, 2017.
- [9] J. Ladley, *Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program*. Waltham, Mass: Morgan Kaufmann, 2012.
- [10] T. C. Redman, *Data Quality for the Information Age*, 1st ed. Norwood, MA, USA: Artech House, Inc., 1997.
- [11] D. Loshin, *Big data analytics: from strategic planning to enterprise integration with tools, techniques, NoSQL, and graph*. Elsevier, 2013.
- [12] M. Rodríguez, J. R. Oviedo, y M. Piattini, «Evaluation of Software Product Functional Suitability: A Case Study», *Softw. Qual. Prof.*, vol. 18, n.º 3, pp. 18-29, 2016.
- [13] I. Caballero, M. Rodríguez, y C. M. Fernández Sánchez, «Calidad de datos digitales certificada», *Rev. AENOR*, p. 4, 2018.
- [14] I. Caballero, A. Gómez, F. Gualo, J. Merino, B. Rivas, y M. Piattini, *Calidad de Datos*. Ra-Ma, 2018.
- [15] ISO/IEC, «ISO/IEC 25012:2008 Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model», ISO/IEC, International Standard, 2008.
- [16] ISO/IEC, «ISO/IEC 25024:2015 Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- Measurement of data quality», ISO/IEC, International Standard, 2015.
- [17] ISO/IEC, «ISO/IEC 25040:2011 Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- Evaluation process», ISO/IEC, International Standard, 2011.

6. AGRADECIMIENTOS

Esta publicación es parte del proyecto de I+D+i PID2020-112540RB-C42, AETHER-UCLM (Una Aproximación Holística de Smart Data para el Análisis de Datos Guiado por el Contexto Centrada en la Calidad y la Seguridad), financiado por MCIN/AEI/10.13039/501100011033/. Esta publicación es parte del proyecto de I+D+i SBPLY/21/180501/000061, ADAGIO, (Alarcos' DAta Governance framework and systems generatIOn) financiado por la Consejería de Educación, Cultura y Deportes de la Junta de Comunidades de Castilla-La Mancha. Esta investigación también ha sido cofinanciada por el Programa de Doctorado Industrial (Ref.: DIN2018-009705) del Ministerio de Ciencia, Innovación y Universidades