

Evaluación de la calidad de datos en un Sistema de Gestión Académica de una Universidad Peruana basado en el estándar ISO/IEC 25000

Marleny PERALTA ASCUE

Escuela Académico Profesional de Ingeniería Informática y Sistemas, Universidad Nacional Micaela Bastidas de Apurímac
Escuela Profesional de Ingeniería de Sistemas e Informática, Universidad Tecnológica de los Andes
Abancay, Apurímac 03001/Apurímac, Perú

José Luis MERMA ARONI

Escuela Académico Profesional de Ingeniería Informática y Sistemas
Universidad Nacional Micaela Bastidas de Apurímac
Abancay, Apurímac 03001/Apurímac, Perú

Carolina SOTO CARRION

Escuela Profesional de Ingeniería Ambiental y Recursos Naturales
Universidad Tecnológica de los Andes
Abancay, Apurímac 03001/Apurímac, Perú

Wilber JIMENEZ MENDOZA

Escuela Académico Profesional de Educación Inicial Intercultural Bilingüe Primera y Segunda Infancia
Universidad Nacional Micaela Bastidas de Apurímac
Abancay, Apurímac 03001/Apurímac, Perú

RESUMEN

Es muy común que las organizaciones desconozcan el nivel de calidad de sus datos almacenados y el valor que estos datos pueden tener para generar grandes beneficios económicos, sociales y científicos. Por lo que, hoy en día resulta imprescindible para cualquier institución disponer de datos de calidad para una buena toma de decisiones y analítica de datos. En este contexto, el presente artículo tiene como objetivo evaluar la calidad de datos sobre la base principal de la familia ISO/IEC 25000, conocida como SQuaRE: *Software Product Quality Requirements and Evaluation*, y el estándar ISO/IEC 25012 como modelo de calidad de datos, aplicada a una base de datos del Sistema de Gestión Académica de la Universidad Nacional Micaela Bastidas de Apurímac-Perú. Las características de calidad de datos fueron seleccionadas desde la perspectiva del consumidor de datos y desde la visión de la calidad de datos inherentes, las métricas fueron calculadas a través de plantillas de consultas SQL.

Palabras Claves: Evaluar, calidad de datos, SQuaRE, Sistema de gestión académica, métricas.

1. INTRODUCCIÓN

Los datos son los activos más reconocidos en el mundo digital y son calificados como activos digitales valiosos para las organizaciones, que requiere una atención muy significativa por parte de los ejecutivos [1][2]. Las organizaciones de hoy deben de adaptarse a una nueva realidad donde los datos van creciendo de manera

exponencial. Esta nueva realidad se define como la “revolución de los datos”, que se conceptualiza como “una explosión en el volumen de los datos, en la velocidad en que los datos son producidos, el número de productores de datos, la difusión de los datos, y el rango de cosas sobre las que hay datos, que provienen de nuevas tecnologías como los celulares e Internet, y de otras fuentes tales como los datos generados por los ciudadanos” [3].

Estudios realizados por [4] señalan que del 100% de los datos estructurados en una organización, menos del 50% son utilizados de manera activa para tomar decisiones, así como, más del 70% de los trabajadores tienen disponibilidad de datos que no deberían y que el 80% del tiempo de los analistas se dedican solo a capturar y preparar datos.

La compañía multinacional *PowerData* de origen español que proporciona servicios especializados de gestión de datos con presencia en países como Argentina, Chile, Colombia, Ecuador, México, Perú y Uruguay, aplicó una encuesta realizada por *PriceWaterhouseCoopers*, en el año 2018, cuyos resultados fueron que el 75% de 600 empresas a quienes encuestaron reportaron problemas significativos como resultado de datos incorrectos, debido a la mala calidad de datos.

De acuerdo a la revisión bibliográfica, existen diferentes concepciones de calidad de datos de las cuales se destacan tres conceptos porque consideran las características más relevantes como son: apropiados para el uso, cumplen con los requisitos de datos y son

consistentes con el mundo real, porque representan la abstracción mundo real [5][6][7][8][9][10].

Evaluar significa “evaluar o estimar la naturaleza, capacidad, o calidad” de alguna una cosa. La calidad de los datos se puede evaluar en diferentes grados utilizando herramientas y técnicas, que van desde la evaluación cualitativa simple hasta la evaluación cuantitativa. El propósito general de la evaluación de la calidad de los datos consiste en comprender el estado actual de los datos y sacar una conclusión sobre su impacto en relación con expectativas o propósitos particulares. Los resultados de la evaluación se pueden emplear para formular o confirmar expectativas relacionadas con la calidad [11].

La presente investigación tuvo como contexto de estudio al Sistema de Gestión Académica (SIGSA) de la Universidad Nacional Micaela Bastidas de Apurímac (UNAMBA), el mismo que se encarga de procesar las matrículas vía intranet e internet del semestre académico correspondiente dentro del calendario académico vigente [12]. En el panorama actual de la universidad, se aprecia que se ve enfrentado a datos inexplorados que pueden tener cierta relevancia y ser un potencial para crear valor. Sin embargo, al explorar y usar los datos del SIGSA para generar información analítica mediante nuevas aplicaciones se presentaron los siguientes problemas:

- Información incompleta; como campos de datos que no se cargaron o están vacíos, cuando son datos obligatorios.
- Valores fuera de un rango referencial.
- Errores de sintaxis, que están directamente relacionados con el formato de datos.
- Datos inconsistentes, datos duplicados o incluso representaciones no adecuadas a la realidad.
- Problemas de estandarización.
- Valores desactualizados.
- Restricciones de integridad no definidas

Por otro lado, el problema general que debilita la gestión académica en la UNAMBA, es la vulnerabilidad crítica efecto de la calidad de datos en la implementación de nuevas soluciones informáticas como: *Business Intelligence, Data Mining, Machine Learning, Big Data, Data Science, Learning Analytics*. De lo expuesto y la relevancia de los datos en el mundo digital, es que surge la necesidad de conocer la magnitud real de la calidad de datos mediante la evaluación.

2. ISO/IEC 25000

El estándar ISO/IEC 25000 o conocido como el estándar SQuARE (*System and Software Quality Requirements and Evaluation*) es un conjunto de normas internacionales compuesta por cinco divisiones. Uno de estos corresponde a la familia 2501n, que se focaliza en modelos de calidad de productos de software y, como

parte de esta familia, se desprende el estándar ISO/IEC 25040 que propone un modelo de evaluación, ISO/IEC 25012 que especifica un modelo general de calidad de datos e ISO/IEC 25024 en donde se definen métricas para realizar la medición de la calidad de datos [13].

Modelo de calidad de datos ISO/IEC 25012

El estándar ISO/IEC 25012, propone un modelo de calidad de datos enfocado a Sistemas de Información en un entorno particular; identifica quince características de calidad que se pueden evaluar desde dos perspectivas: *inherente y dependiente del sistema*.

- **Inherente.** Hace referencia al grado en el cual las características de calidad del dato dependen de su propia naturaleza, cubren necesidades implicadas bajo condiciones específicas [14].
- **Dependiente del sistema.** Hace referencia al grado en el cual el nivel de calidad del dato depende del sistema de información donde los datos se encuentran almacenados y del dominio tecnológico en el que se hace uso del dato [14].

En la Tabla 1 se muestran las características de la calidad de datos de acuerdo al estándar ISO/IEC 25012 [14].

Tabla 1. Características de calidad de datos según el estándar ISO/IEC 25012 [14].

Perspectiva	Características
Inherentes	Exactitud Compleitud Credibilidad Actualidad
Inherentes y dependientes del sistema	Accesibilidad Conformidad Confidencialidad Eficiencia Precisión Trazabilidad Entendibilidad
Dependientes del sistema	Disponibilidad Portabilidad Recuperabilidad

Medidas de calidad de datos: ISO/IEC 25024

El estándar ISO/IEC 25024, propone un conjunto de medidas y métricas que incluyen la forma de medir las características de calidad de datos de ISO/IEC 25012. Cuantifican la magnitud en que un conjunto de datos cumple con ciertos valores [15]. Cada característica de calidad de datos tiene uno o más elementos de medición que pueden ser utilizados para medir la calidad del valor del dato y la calidad del modelo conceptual de datos [16].

Proceso de evaluación: ISO/IEC 25040

El estándar ISO/IEC 25040, “proporciona una descripción del proceso de evaluación de la calidad de un producto software y establece los requisitos para la aplicación de este proceso” [17]. También puede ser usada para evaluar la calidad de un repositorio de datos. El proceso de evaluación tiene cinco actividades como se visualiza en la Figura 1.

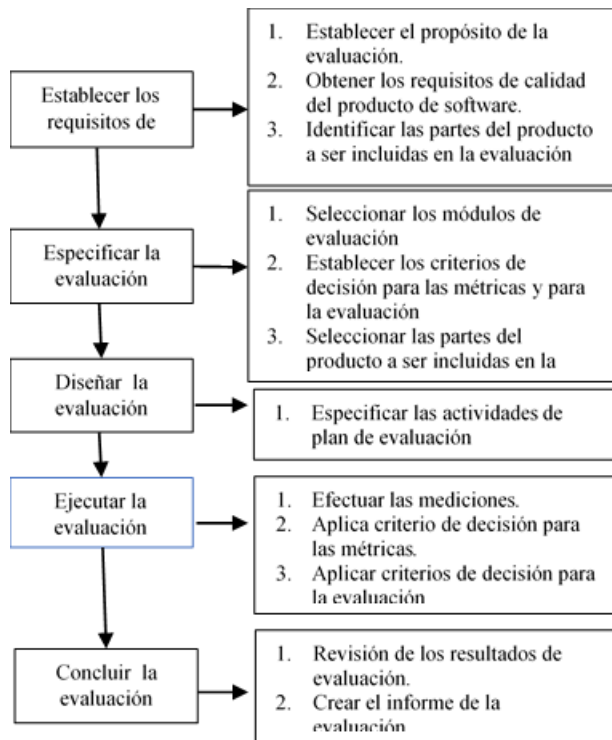


Figura 1. Proceso de evaluación de calidad ISO/IEC 25040 [18]

3. MATERIALES Y MÉTODOS

La muestra está constituida por los datos almacenados en el SISGA desde el año 2010 al 2017, con un total de 22 tablas, 154 atributos y 319,685 registros. Datos que fueron proporcionados por el responsable de la Dirección de Servicios Académicos de la UNAMBA para la ejecución de proyectos de investigación. La investigación se realiza respetando la Ley de protección de datos personales, publicada en el año 2011, por el Gobierno del Perú [19].

El proceso de evaluación se realizó siguiendo cada una de las etapas establecidas en ISO/IEC 25040.

Método de selección de las características de calidad según ISO/IEC 25012

En la etapa de la especificación de la evaluación, se selecciona las características de calidad, métricas y criterio de decisión para la evaluación. La selección se

realizó desde la perspectiva del consumidor de datos, porque son las personas que hacen usos de los datos para fines analíticos, cualitativos, cuantitativos o predictivos, para proyectar reportes e indicadores, en ese sentido necesitan usar los datos, cuyos valores se espera que sean correctos, actualizados y completos; y desde la perspectiva de calidad de datos inherente para que la evaluación de los datos vaya más allá de aspectos tecnológicos.

Se aplicó una encuesta a los expertos de sistemas académicos para determinar la importancia de las características de calidad de datos. Considerando las valoraciones de los expertos ingenieros responsables de Sistemas de Gestión Académica de dos universidades de Apurímac, se empleó el Método de Jim Brosseau para incluir el criterio del investigador (evaluador) en la selección de características y métricas apropiadas [20].

Brosseau, plantea un método validado que es utilizado para cuantificar y priorizar atributos o características de calidad en un contexto específico. El método está compuesto de dos etapas:

1. Se identifica los atributos de calidad que estén dentro del alcance de la evaluación y que son aplicables al caso de estudio, a partir de ello se comenzará a filtrar las características.
2. Priorización de atributos, en función a los resultados del ítem anterior, se elabora una matriz de características restantes para realizar una comparación por pares y se identifica cuál es más importante en cada par, luego se considera el puntaje obtenido y en base a ello se prioriza los atributos.

Finalmente, se decidió establecer el conjunto de las características para medir la calidad de datos: Exactitud, Completitud, Consistencia, Credibilidad, Actualidad, Conformidad, Precisión y Extensibilidad. En la Tabla 2, se muestra la definición de las características [21] y en la Tabla 3 las métricas por cada característica.

Tabla 2. Definición de las características de calidad de datos seleccionadas según el estándar ISO/IEC 25012 [14]

Característica	Definición
Exactitud	Se refiere al grado en el cual el dato tiene atributos que representan correctamente el valor verdadero del atributo esperado de un concepto o evento en un contexto específico de empleo.
Completitud	Se refiere al grado en el cual el dato del sujeto asociado con una entidad tiene valores para todos los atributos esperados e instancias de entidad relacionadas en un contexto específico de uso.

Consistencia	Se refiere al grado en el cual el dato tiene los atributos que son libres de contradicción y son coherentes con otros datos en un contexto específico de uso.
Credibilidad	Se refiere al grado en el cual el dato tiene atributos que son considerados como verdaderos y auténticos en un contexto específico de uso.
Actualidad	Se refiere al grado en el cual el dato tiene los atributos que son correctos en un período de contexto específico de uso.
Conformidad	Se refiere al grado en el cual el dato tiene atributos que se adhieren a normas, convenciones o regulaciones vigentes y reglas similares relacionadas con la calidad de datos en un contexto específico de uso.
Precisión	Se refiere al grado en el que el dato tiene atributos que son exactos o que proporcionan la discriminación en un contexto específico de uso.
Entendibilidad	Se refiere al grado en el cual el dato tiene atributos que le permiten ser leído e interpretado por usuarios, y es expresado en lenguajes apropiados, símbolos y unidades en un contexto específico de uso.

Tabla 3. Características y métricas de calidad de datos seleccionadas ISO/IEC 25024 [22]

Características	Código	Descripción de la métrica
Exactitud	ACC-I-1	Exactitud sintáctica de los datos
	ACC-I-2	Exactitud semántica de los datos
	ACC-I-3	Aseguramiento de la exactitud de los datos
	ACC-I-7	Rango de Exactitud de datos
Complejidad	COM-I-1	Complejidad de registros
	COM-I-2	Complejidad de atributos
Consistencia	CON-I-2	Consistencia del formato de datos
	CON-I-3	Riesgo de inconsistencia de datos
	CON-I-5	Cobertura de la consistencia de valores de datos
	CON-I-6	Consistencia semántica
Credibilidad	CRED-I-1	Credibilidad de los valores
Actualidad	CUR-I-2	Oportunidad de la actualización
	CUR-I-3	Requisito de actualización de ítems

Conformidad	CNF-I-1	Conformidad regulatoria de los valores y/o formato de datos
Precisión	PRE-I-1	Precisión de valores de datos
Entendibilidad	UND-I-4	Entendibilidad de valores de datos

Para definir el criterio de decisión (niveles de aceptabilidad), se ha considerado las siguientes evaluaciones objetivas: Razón o relación simple que mide la relación entre los resultados deseados y los resultados totales, por otro lado, se utilizó la operación de mínimos y máximos para algunas métricas como credibilidad y oportunidad. La escala de valores para el proceso de evaluación va de 0 a 1, por lo que de 0 a 0,74 la calidad de datos es inaceptable, mientras que de 0,75 a 1.00 la calidad de datos es aceptable, cuanto más cercano a uno mejor [23].

Para realizar el cálculo de las métricas, se elaboró plantillas de consultas SQL (*Structured Query Language*).

4. RESULTADOS

Los resultados de la evaluación de los datos del Sistema SIGSA se muestran en la Tabla 4.

Tabla 4. Resultados de la evaluación

Característica	Código de la métrica	Resultado	Resultado global.	Criterio de decisión
Exactitud	ACC-I-1	0,82	0,82	Aceptable
	ACC-I-2	0,75		
	ACC-I-3	1,00		
	ACC-I-7	0,69		
Complejidad	COM-I-1	0,74	0,77	Aceptable
	COM-I-2	0,80		
Consistencia	CON-I-2	0,39	0,77	Aceptable
	CON-I-3	0,99		
	CON-I-5	1,00		
	CON-I-6	0,70		
Credibilidad	CRED-I-1	0,78	0,78	Aceptable
Actualidad	CUR-I-2	0,69	0,56	Inaceptable
	CUR-I-3	0,42		
Conformidad	CNF-I-1	0,56	0,56	Inaceptable
Precisión	PRE-I-1	0,74	0,74	Inaceptable
Entendibilidad	UND-I-4	0,73	0,73	Inaceptable
Promedio			0,72	

Los resultados muestran en promedio un 72% de calidad de datos, por lo que se puede decir que se encuentra en niveles no aceptables, siendo las métricas de actualidad, conformidad, precisión y entendibilidad las que tienen menor porcentaje de calidad. Se aprecia Esto se debe a que los datos no son actualizados, no cumplen las reglas de actualización. Sobre la conformidad de datos, no tienen formatos adecuados, convenciones o normativas vigentes, por consecuencia no se tiene valores precisos. En cuanto a la característica de entendibilidad existen valores en los datos con formatos no entendibles como por ejemplo en el atributo sexo considerado como 1 y 2, cuando debería ser M o F. En la Figura 2, se muestra el gráfico radar de los resultados promedio por cada característica.

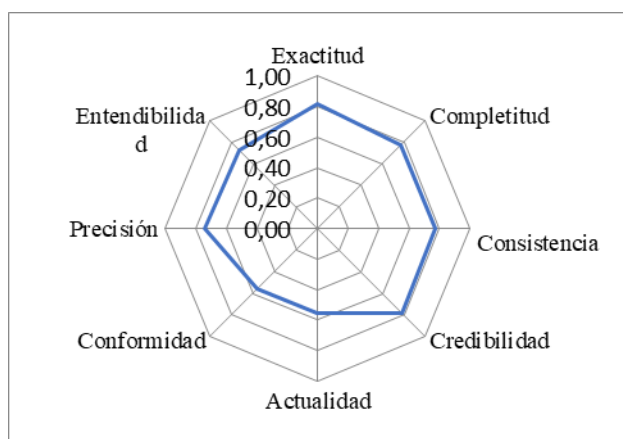


Figura 2. Gráfico radar de los resultados de cada característica de calidad de datos.

Analizando los resultados de las características con un bajo nivel de calidad a nivel de tablas, se puede obtener análisis importantes de los registros catalogados como datos de mala calidad en el Sistema SISGA, como, por ejemplo: en los datos de los docentes y estudiantes con porcentajes bajos de calidad de datos. El 61% de datos de los docentes y 63% de los estudiantes no disponen de información exacta, el 31% de los datos de estudiantes y docentes no son completos, son inconsistentes y no estandarizados.

5. CONCLUSIONES

Si bien la universidad cuenta con un Sistema de Gestión Académica considerado como fortaleza dentro de la institución, esto se debilita al no contar con datos de calidad y esto se evidencia al evaluar la calidad de datos con características de calidad con un bajo nivel de aceptabilidad. En consecuencia, la universidad en tema de calidad de datos no le dio la debida importancia.

En todo proceso de evaluación de calidad de datos, es necesario trabajar de la mano con la especificación de requisitos, esto debido a que ciertas métricas de calidad lo requieren como insumo para su medición.

El contar con datos de calidad se hace necesario para realizar una buena analítica de datos. Sin datos no se puede generar información, por lo que resultaría imposible crear conocimiento.

6. REFERENCIAS

- [1] Westerman, G., Bonnet, D. & McAfee, A. (2014) *Leading digital: Turning technology into business transformation*. Boston, United State: Harvard Business Press
- [2] Piattini, M., Gómez, A., Caballero, I., Gualo, F., Merino, J y Rivas, B. (2018). *Calidad de Datos*. Madrid, España: Editorial RAMA.
- [3] High-Level Panel of Eminent Persons on the Post-2015 Development Agenda, "A New Global Partnership: Eradicate Poverty and Transform Economies through Sustainable Development," 2013.
- [4] DalleMule, L. & Davenport T. (2017). *What's your Data Strategy*. United States: Harvard Business. Recuperado de: <https://hbr.org/2017/05/whats-your-data-strategy>.
- [5] Wand, Y, & Wang, R. (1996) Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86-95.
- [6] Redman, T. (Febrero, 2013). *Data Quality Management Past, Present, and Future: Towards a Management System for Data*. In S. Sadiq (Ed.), *Handbook of Data Quality SE - 2* (pp. 15-40). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36257-6_2
- [7] Manrique, A. (2017). Desarrollo de un catálogo de reglas de negocio referentes a datos, basado en ISO/IEC 25012 y SBVR. (Tesis de Magister). Universidad de Castilla-La Mancha, España.
- [8] Colin, L. (2007). Las normas ISO 9000:2000 de sistemas de gestión de la calidad; *Revista Virtual Pro, Sistemas de Gestión de calidad*. Recuperado de: <https://www.ineel.mx/bolISO02/tecni2.pdf>
- [9] Wand, Y, & Wang, R. (1996) Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86-95.
- [10] Pio, S. (2018). *Minería de calidad de datos: Aplicación de Técnicas de minería de datos, para la evaluación de la calidad de datos*. (Tesis de Magister). Universidad de la República, Montevideo, Uruguay.
- [11] Coleman, L. (2013). *Measuring Data Quality for Ongoing Improvement. A Data Quality Assessment Framework*. United States: Morgan Kaufmann.

- [12] UNAMBA (2018). *Reglamento de matrícula*. Universidad Nacional Micaela Bastidas de Apurímac, Perú.
- [13] ISO/IEC 25000. (2021). System and software engineering – Systems and software product Quality Requirements and Evaluation (SQuaRE) – Evaluation Process, International Organization for Standardization. Geneva, Switzerland.
- [14] ISO/IEC 25012. (2008). Software engineering – Software product Quality Requirements and Evaluation (SQuaRE)– Data Quality
- [15] Piattini , M., Gómez, A., Caballero, I., Gualo, F., Merino, J y Rivas, B. (2018). *Calidad de Datos*. Madrid, España: Editorial RAMA
- [16] ISO/IEC 25024. (2015). Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of data quality.
- [17] NTE INEN-ISO/IEC 25040 (2014). Sistemas e Ingeniería de software- Requerimientos y evaluación de sistemas de calidad de software (SQuaRE) – Proceso de Evaluación. Instituto ecuatoriano de Normalización. Quito, Ecuador.
- [18] ISO/IEC 25040. (2005). System and software engineering – Systems and software product Quality Requirements and Evaluation (SQuaRE) – Evaluation Process, International Organization for Standardization. Geneva, Switzerland.
- [19] Congreso de la República de Perú (2011) Ley 29733, *Ley de Protección de Datos Personales*. http://www.pcm.gob.pe/transparencia/Resol_ministeriales/2011/ley-29733.pdf
- [20] Brosseau, J. (2010). Software quality attributes: Following all the steps. Clarrus Consulting Group Inc.
- [21] Guerra, C. A. (2011). *Entorno de Trabajo para la Gestión de Requisitos de Calidad de Datos en el Desarrollo de Aplicaciones Web*. (Tesis Doctoral). Universidad de Castilla- La Mancha. España.
- [22] Fernández, M., Dávila, A. & Angeleri, P. (Octubre, 2017). Data quality applied to an academic business intelligence solution: Lesson learned. *IEEE Colombian Conference on Communications and Computing (COLCOM)*, Cartagena, Colombia, pp. 1-6, doi: 10.1109/ColComCon.2017.8088200.
- [23] Calahorrano, A.I. Herramienta para evaluar la calidad del código fuente generado en C ANSI. (Tesis Título Profesional). Escuela Politécnica Nacional. Ecuador.