Universidad de Medellín
Ciencia y Libertad
Resolución 103 del 31 de julio de 1950

El Doctorado en Modelación y Computación Científica
de la Facultad de Ciencias Básicas

informa que el

Dr. Diego Hernán Peluffo Ordoñez
Mohammed VI Polytechnic University

participó en la dirección de tesis doctoral del estudiante **Miguel Alberto Becerra Botero** titulada: "Framework de fusión de datos y calidad de la información para sistemas de información".

Dado en Medellín, a los 30 días del mes de mayo de 2023.

JOSÉ ALBERTO RÚA VÁSQUEZ

Dr. José Alberto Rúa Vásquez
Decano
Facultad de Ciencia Básica

Dr. Francisco Caro
Coordinador Doctorado en
Modelación y Computación Científica

FRAMEWORK DE FUSIÓN DE DATOS Y CALIDAD DE LA INFORMACIÓN PARA SISTEMAS DE INFORMACIÓN

Presentada por:
Miguel Alberto Becerra Botero

Tesis Doctoral por compendio de artículos, para optar al título de:
Doctor en Modelación y Computación Científica

Dirigida Por:
Diego Peluffo Ordóñez Ph.D.
Catalina Tobón Zuluaga Ph.D.

Universidad de Medellín
Facultad de Ciencias Básicas
Medellín, Colombia
2023

AGRADECIMIENTOS

Agradezco a Dios, a la Ph.D. Catalina Tobón y al Ph.D. Diego Peluffo por su apoyo incondicional.

TABLA DE CONTENIDO

RESUMEN

La fusión de la información tiene como objetivo mezclar información para mejorar la completitud y precisión, respecto a información obtenida de fuentes individuales y su desempeño, sin embargo, está limitada principalmente a la valoración de estos dos criterios de calidad. Adicionalmente, la calidad de la información es medida en la mayoría de los sistemas de información por múltiples dimensiones o criterios dependientes de los requerimientos de usuario y del proceso. La fusión de la información y la calidad de la información son campos abiertos de investigación que han sido poco estudiados en conjunto.

Por otro lado, los modelos y arquitecturas de fusión de datos reportados en la literatura son bastante amplios, siendo el modelo *Joint Directors of Laboratories* (JDL por sus siglas en inglés) uno de los modelos funcionales más populares que cubre desde el proceso de adquisición hasta el proceso de valoración de la situación y del riesgo. Sin embargo, la valoración de la calidad de la información ha sido poco explorada como parte integral de este modelo funcional con el fin de realizar procesamientos que permitan optimizar la calidad de la información entregada al usuario final, de acuerdo con los requerimientos demandados por este y siguiendo escalonadamente la cadena de procesamiento.

En esta tesis por compendio de artículos, se propuso un *framework* de propósito general que permite construir sistemas de fusión de datos en el marco del modelo JDL cubriendo los niveles funcionales que demanda la aplicación, como son, fusión de bajo y/o alto nivel con capacidad de optimización considerando múltiples criterios de calidad para cada nivel y evaluando dependencias entre ellas para definir los criterios de calidad que deben ser mejorados para cumplir los requerimientos del usuario final y así entregar no sólo un procesado y fusión de los datos útil para la toma de decisiones, sino también la valoración de calidad de la información, tanto a nivel local de cada nivel en el marco del modelo JDL, como a nivel global.

Hasta la fecha, este enfoque no ha sido integrado a este modelo en un *framework* conjunto. La propuesta incluye un análisis teórico del diseño de la arquitectura de propósito general y la descripción de cada uno de sus procesos funcionales junto con su respectiva metodología de aplicación (fase teórica). En la fase experimental se aplica la propuesta en diferentes ambientes de prueba, como son: sistema de variables ambientales y contaminantes, sistema enfocado al procesado de electrogramas, un sistema enfocado a la biometría y un sistema enfocado al capital intelectual. Los resultados experimentales demuestran la funcionalidad de la propuesta en múltiples ambientes y la importancia del uso de la calidad de la información como parte integral de los procesos funcionales y de refinamiento en los sistemas de fusión de datos para su optimización y trazabilidad. De igual forma, los resultados permitieron establecer limitaciones que deben ser afrontadas en trabajos futuros como desafíos y realizar pruebas en otros ambientes para ampliar la generalidad y funcionalidad del *framework* propuesto.

ABSTRACT

Information fusion is aimed at merging information to improve the quality thereof, being expected to outperform results from individual sources, in terms of two quality criteria: completeness and accuracy. Its performance is mainly limited to the assessment of these two quality criteria. Information quality is measured in most information systems by multiple dimensions or criteria dependent on user and process requirements. Information fusion and information quality are open fields of research and have not been widely studied in a joint manner.

The data fusion models and architectures reported in the literature are quite extensive, with the Joint Directors of Laboratories (JDL) model being one of the most popular functional models, covering from the acquisition process to the situation and risk assessment process. However, prior to our review of the literature, no studies were found that have involved the information quality assessment applied on this functional model considering multiple criteria that allow processing to optimize the information quality delivered to the end user according to the requirements demanded by the end user and following the processing chain in a staggered manner.

In this thesis, presented in the article compendium modality, a general-purpose framework is proposed that allows building data fusion systems within the JDL model framework, covering the functional levels demanded by the application, for example, low and/or high-level fusion with optimization capabilities, considering multiple quality criteria for each level and evaluating dependencies between them, in order to define which quality criteria should be improved to meet the end-user requirements and thus deliver not only a data processing and fusion useful for decision making, but also the information quality assessment both at the local level in the JDL model framework together a global assessment.

To date, this approach has not been integrated into this model in a joint framework. The proposal includes a theoretical analysis of the general-purpose architecture design and the description of each of its functional processes together with their respective implementation methodology. In the experimental phase, the proposal is applied to different test environments, such as: a system of environmental variables and pollutants, a system focused on electrogram processing, and a system focused on biometrics. The experimental results demonstrate the functionality of the proposal in multiple environments and the importance of the use of information quality as an integral part of the functional and refinement processes in data fusion systems for their optimization and traceability. Likewise, the results allowed establishing limitations that must be faced in future works and the need to perform tests in other environments to extend the generality and functionality of the proposed framework.

1. INTRODUCCIÓN

1.1 CONCEPTOS BÁSICOS

A continuación, se exponen conceptos básicos y fundamentos de los sistemas de fusión de datos y calidad de la información, necesarios para comprender la relación de la fusión de datos en función de la calidad de la información como parte fundamental en el desarrollo de este trabajo.

Fusión de datos

Términos como fusión de datos y fusión de la información, han sido utilizados indistintamente en la literatura (Abdelgawad & Bayoumi, 2012). A pesar de que describen una misma tarea, estos términos tienen algunas variantes respecto a la aplicación y tipo de datos, que puede ser difícil diferenciar. Algunos estudios cuando hacen referencia a la fusión de datos, se refieren a datos tomados directamente de los sensores sin ningún tipo de procesamiento, mientras que la fusión de la información es considerada como la fusión de datos que ha sufrido previamente algún tipo de procesamiento (Castanedo, 2013). En las ciencias de la computación, la diferencia entre dato e información radica en el contexto y en el significado (Rein & Biermann, 2013).

Fusión de datos tiene un concepto bastante difuso que toma diferentes interpretaciones con las aplicaciones y fines específicos. En la literatura se han reportado varias definiciones que hace difícil discernir entre fusión de la información y fusión de datos. La definición de fusión de datos más acogida es la del modelo *Joint Directors of Laboratories* (JDL por sus siglas en inglés). Sin embargo, en este documento y para dar una mayor generalidad, se aceptan las siguientes definiciones reportadas en la literatura (Boström et al., 2007): i) Conjunto de metodologías y tecnologías que posibilitan la combinación sinérgica de datos heterogéneos de distintas fuentes en conjunto con nuevos datos, conteniendo mayor información que la suma de las fuentes individuales. ii) Proceso multinivel para el manejo de procesos multifacéticos de la detección automática, de asociación, correlación, estimación y combinación de datos e información de varias fuentes. iii) Proceso de varios niveles frente a la asociación, correlación, combinación de datos e información de fuentes únicas y múltiples para lograr un estado refinado o íntegro completo, un reconocimiento de estimaciones y evaluaciones completas y oportunas de las situaciones de amenazas y su importancia.

Aunque se han propuesto múltiples modelos de fusión de datos, el modelo JDL es uno de los más utilizados por la comunidad de fusión de datos, es considerado el modelo más popular y ha sido ampliamente utilizado como guía para el diseño de sistemas de fusión de la información. Por lo anterior, el presente trabajo está basado en dicho modelo. El modelo JDL fue propuesto como un modelo de cuatro niveles (White, 1991) y posteriormente ajustado a cinco niveles (Steinberg et al., 1999). Este consiste de un bus de datos que conecta cinco niveles de procesamiento como se ilustra en la Figura 1.

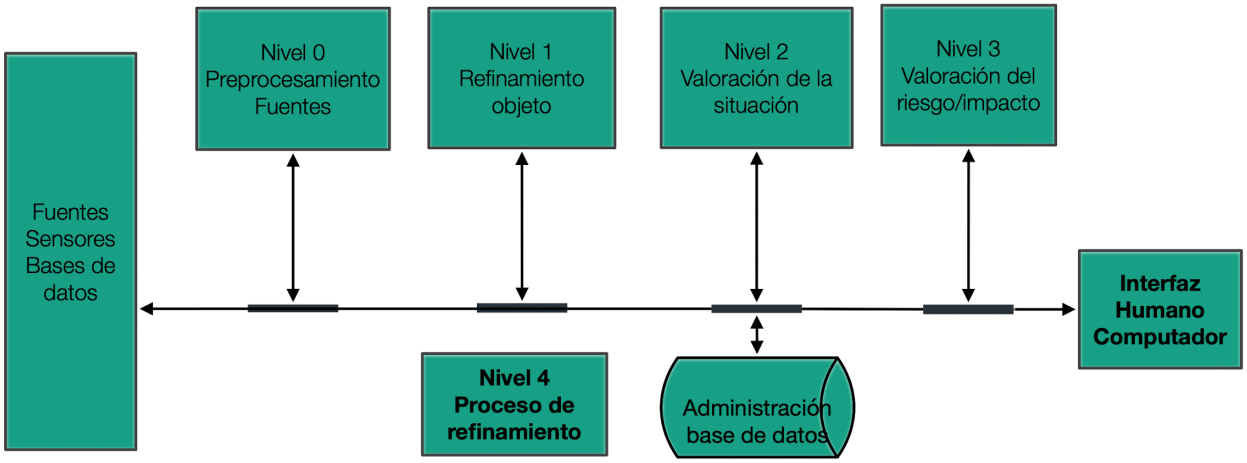

Figura 1. Arquitectura JDL.

Para realizar la fusión de datos se hace uso de diferentes técnicas entre las que se encuentran las probabilísticas, *soft-computing*, algoritmos de optimización, entre otros; cuyo uso (caracterización, estimación, agregación, clasificación, compresión, entre otros) depende del tipo de aplicación. Adicionalmente, se deben considerar las ventajas y desventajas de las técnicas para una adecuada selección, con el fin de obtener un efectivo desempeño. En la Tabla 1 se presentan las principales características de diferentes metodologías usadas en fusión de datos, resaltando sus ventajas y desventajas, lo cual es considerado en la etapa de selección de los algoritmos de fusión de datos cuando se realiza el proceso de ingeniería para la construcción de un sistema de fusión de datos.

Tabla 1. Metodologías de fusión de datos

	Categoría	Características	Ventajas	Desventajas
Técnicas de asociación de datos (Castanedo, 2013)	Probabilística	Formulado como problema de inferencia bayesiana, resuelto usando filtros de Kalman o filtros de partículas.	Bien establecido e investigado, enfocado a determinar la incertidumbre.	Alta complejidad, inconsistente, baja precisión del modelo, incertidumbre acerca de la incertidumbre.
	Asociación de datos probabilísticos (PDA)	El algoritmo asigna una probabilidad de asociación de cada hipótesis desde una medida válida de un *target (entidad objetivo)* y la estimación del estado puede ser realizada usando una suma ponderada de los estados estimados bajo todas las hipótesis.	El algoritmo puede asociar diferentes medidas a un *target* específico y ayuda a estimar el estado del *target*.	Perdida de *tracks* (rastreo de entidades) ya que no contempla otros *targets* cercanos. Problemas para el manejo de la incertidumbre (aproximación bayesiana sub-óptima). Bajo desempeño cuando hay múltiples *targets*. Mal funcionamiento para detección falsas alarmas.
	Asociación de datos probabilísticos conjunta (JPDA) Asociación de datos probabilístico conjunto distribuido (JPDA-D)	A diferencia de PDA las probabilidades de asociación son computadas usando todas las observaciones y todos los *targets.*	Adecuados en ambientes donde se presentan altos niveles de medidas falsas.	Bajo desempeño en rastreo de *targets* en ambientes desordenados. Dos medidas no pueden ser originadas de un mismo *target* en un mismo instante de tiempo. Requieren mecanismos explícitos para inicializar el *track*. No pueden inicializar *tracks* que están fuera del área de observación. Alto costo computacional con incremento exponencial con el

				número de *targets*.
	Test de Hipótesis múltiples (MHT) MHT Distribuido	Combina la asociación de datos y el rastreo en un solo *framework*. Utiliza Bayes para calcular las hipótesis.	Adecuado desempeño en rastreo de *targets* en ambientes desordenados.	Alto costo computacional.
	MHT probabilístico (PMHT)	Considera las asociaciones como variables aleatorias para considerar independencia.	Ventajas del MHT distribuido.	Reduce el costo computacional respecto a MHT.
	Modelos Gráficos: Redes Bayesianas (gráfico directo) Campos aleatorios de Markov (Grafico indirecto)	Representan una descomposición condicional de la probabilidad conjunta.	Útiles para representar relaciones causales entre variables aleatorias. El gráfico indirecto es útil para expresar restricciones *soft* entre variables aleatorias. Bayes asume independencia.	Limitados a resolver problemas de asociación de datos distribuidos en redes de sensores sincronizadas con áreas traslapadas y donde cada sensor recibe ruido.
	Vecinos más cercanos	Selecciona o agrupa los valores similares y está basado en una distancia métrica (estadística, euclidiana o absoluta) y un umbral.	Fácil solución con mínimo tiempo de procesamiento.	Propagación de errores (pares con la misma probabilidad). Bajo desempeño en ambiente con ruido.
	K-Means	Divide los datos en conjuntos dentro de diferentes clústers.	Simplicidad. K-means ponderado reduce las iteraciones cuando se requiere determinar el número de clúster óptimo.	El algoritmo no siempre encuentra la solución óptima y el número de clúster debe ser conocido a priori y se asume como óptimo. Considera la covarianza de los datos como irrelevante o que ha sido normalizada.
Métodos de estimación de estados (Castanedo, 2013)	Máximo posterior (MAP)	Está basado en la teoría bayesiana.	Buen desempeño cuando las variables de estado siguen una distribución de probabilidad desconocida.	Requiere el modelo analítico del sensor para proporcionar la distribución a priori y computar la función de vecindad.
	Máxima vecindad (ML)	Método de estimación basado en la teoría probabilística. Es utilizado cuando solo hay observaciones y no existe información a priori sobre *x*.	Buen desempeño cuando la variable de estado sigue una distribución de probabilidad desconocida. Desestima el problema de varianza de la distribución incrementando el número de datos.	
	Filtro de Kalman Filtro de Kalman Extendido Filtro de Kalman distribuido	Son utilizados principalmente para la fusión de datos de bajo nivel.	Simplicidad, fácil implementación y optimalidad en términos del error cuadrático medio. Kalman extendido es útil para implementar filtros recursivos no lineales (sistemas no lineales). Fácil implementación en paralelo.	Kalman extendido tiene costo computacional muy alto y se vuelve intratable en altas dimensiones. Alta sensibilidad ante datos corruptos con *outliers*.
	Filtro de Partículas Filtro de Partículas distribuido	Recursiva implementación del método de Monte Carlo. Ante altas dimensiones el algoritmo es un Modelo Monte Carlo Cadena de Markov (MCMC)	Es un filtro más flexible que el filtro de Kalman. Puede enfrentar dependencias no-lineales y densidades no Gaussianas en el modelo. Modelos recientes usan un número dinámico de partículas lo que reduce el costo computacional. El filtro distribuido puede monitorear un ambiente que podría ser capturado por un	Se requiere un gran número de partículas para obtener una pequeña varianza en el estimador generando un alto costo computacional.

			modelo de espacio de estado markoviano.	
	Métodos de consistencia - covarianza	*Frameworks* tolerantes a fallos para mantener la media de la covarianza y las estimaciones en redes distribuidas.	Mejora la restricción del filtro de Kalman distribuido respecto al conocimiento de las covarianzas.	
Métodos de Fusión de Decisión (Castanedo, 2013)	Teoría de Dempster-Shafer	Generaliza la teoría Bayesiana. Fusiona datos de sensores representados como plausibles y creíbles usando reglas de combinación de evidencias dadas.	Proporciona una generalización a la probabilidad con una muy enriquecida representación de la creencia.	Crece la complejidad exponencialmente con la cardinalidad. Asume independencia.
	Métodos Bayesianos	Fusión de la información basada en la inferencia bayesiana lo que permite combinar evidencia basada en reglas de teoría de la probabilidad.	Facilidad de aplicación.	Problema para establecer las probabilidades a priori. Las hipótesis deben ser mutuamente exclusivas. Dificultad en describir la incertidumbre de las decisiones. Complejidad ante múltiples hipótesis.
	Métodos Semánticos	Emplea los datos semánticos desde diferentes fuentes y consta de dos etapas: construcción del conocimiento y la inferencia.	Utiliza la interpretación semántica entregada por los sensores. Permite el traslado de los datos de los sensores a un lenguaje formal. Bajo costo de transmisión.	Complejidad para el almacenamiento de los datos.
	Razonamiento abductivo: redes neuronales, lógica difusa, y otros métodos de inferencia	Método de razonamiento donde una hipótesis es seleccionada bajo la suposición que el caso es verdadero.	El método intenta encontrar la mejor explicación cuando un evento es observado.	Es más un razonamiento de patrones que una técnica de fusión de datos.
	Soft computing	Despliega razonamiento difuso impreciso para combinar datos de sensores fusificados.	Potente esquema de procesamiento en paralelo, cercano al pensamiento humano.	Dificultad en el procedimiento de entrenamiento.
	Basado en la optimización	Formula el problema de la fusión como una optimización de una función de costo definida heurísticamente.	Fácil de integrar nuevos criterios de comportamiento. Abundancia de métodos de optimización para abordar la fusión.	Problema de extremos locales, restricciones en la optimización pueden ser intratables.
	Enfoques Híbrido	Combina diferentes métodos de fusión dentro de una formulación unificada.	Combina las bondades de diferentes técnicas para alcanzar un comprensivo tratamiento de la incertidumbre de los datos.	Costo computacional debido a que tiene múltiples unidades de fusión.
	Teoría de copulas	Se aplica para modelar funciones marginales de los datos y la estructura de dependencia presente entre ellos, simplificando el problema del modelado de datos heterogéneos y dependientes.	Realiza fusión de datos *soft*, fusión *hard* y asume dependencia o no dependencia estadística de los datos, ya que esto degrada los resultados al realizar la fusión.	Existen diferentes familias de cópulas cuya selección depende del tipo de datos y aplicación.

Calidad de la información

La calidad de la información se centra en el usuario (humano o máquina automatizada) y puede ser definida como "el grado en que la información satisface las necesidades del usuario de acuerdo con las percepciones subjetivas externas del usuario" (Wang & Strong, 1996).

Criterios de la calidad de la información

La calidad de la información ha sido abordada como un problema de múltiples dimensiones, por lo cual la mayoría de los modelos y metodologías reportados, se enfocan en propuestas de criterios para la valoración de la calidad de la información desde diferentes perspectivas. El modelo más popular de valoración de la calidad de la información es el presentado en (Wang & Strong, 1996) donde formularon un conjunto de criterios agrupados en 4 categorías así: i) Intrínseco, se refiere a la calidad del dato en sí mismo. ii) Contextual, hace referencia a la valoración hecha en el contexto de uso. iii) Representacional, utiliza criterios enfocados a la presentación de los datos al usuario final. iv) Accesibilidad, se enfoca en evaluar el acceso a los datos y su seguridad. El conjunto de criterios correspondientes a cada grupo es presentado en la Tabla 2.

Tabla 2. Criterios de calidad propuestos por Wang and Strong, 1996

Intrínseco	**Contextual**	**Representacional**	**Accesibilidad**
Exactitud	Valor agregado	Interpretabilidad	Acceso
Credibilidad	Relevancia	Facilidad de entendimiento	Seguridad
Objetividad	Oportunidad	Consistencia en la representación	
Reputación	Completitud	Concisión	
	Data Amount	Manipulabilidad	

Norma ISO/IEC 25012

Esta norma describe un modelo de calidad de datos para evaluar un "producto de datos" (Portal ISO 25000, n.d.) y define calidad como el grado de satisfacción de los requisitos demandados por la organización a la que pertenece el producto. Este modelo cuenta con 15 criterios (ver Tabla 3), los cuales permiten evaluar la calidad de los datos inherente (calidad intrínseca del dato) y la calidad de los datos dependiente del sistema (afectación del sistema sobre los datos). Los criterios que se resaltan en la tabla (columna 2 y 3), corresponden a criterios que pueden ser considerados tanto de la calidad inherente como dependiente del sistema dependiendo de la aplicación.

Tabla 3. Criterios de calidad Norma ISO/IEC 25012

Calidad inherente		**Calidad dependiente del sistema**	
Exactitud	Accesibilidad	Precisión	Disponibilidad
Completitud	Conformidad	Trazabilidad	Portabilidad
Consistencia	Confidencialidad	Comprensibilidad	Recuperabilidad
Credibilidad	Eficiencia		
Actualidad			

Relación entre fusión de datos, calidad de la información y contexto

La fusión de datos y la calidad de la información ha sido analizada en conjunto por (Becerra et al., 2018; Rogova, 2016; I. Todoran et al., 2015). En (Rogova & Snidaro, 2018) fue incluido el contexto y su calidad dentro del análisis, como factor adicional a la calidad de la información en la fusión

de datos. En (I. Todoran et al., 2015) fue propuesta una metodología para evaluar la calidad de la información en los sistemas de información y la llevó a los sistemas de fusión de datos, en la que se realiza una descomposición granular de los sistemas de fusión de datos en subsistemas, permitiendo obtener funciones que los modelan a partir de los datos/información y sus medidas de calidad (entrada/salida), lo cual permite propagar dichas medidas a través de los subsistemas para obtener una medida de calidad global.

Por otro lado, Rogova en su trabajo (Rogova, 2016) ha propuesto un conjunto de criterios de calidad de la información para los sistemas de fusión de datos. En (Rogova & Snidaro, 2018) han considerado la calidad de la información del contexto (características de una situación actual como su información estadística o relaciones entre elementos situacionales bajo consideración), ya que el contexto ha demostrado ser un elemento relevante de información para los procesos de fusión de datos. Adicionalmente, demuestra su aplicación en un escenario de *tracking* considerando el contexto y la calidad de la información como elementos esenciales para la valoración de la calidad de la información. Otra consideración realizada por esta autora en (Rogova, 2016) es la valoración de la calidad de la información, lo cual lleva a un nivel más complejo para la generación de metadatos de los metadatos.

1.2 PLANTEAMIENTO DEL PROBLEMA

La fusión de datos es definida en (White, 1991) como “proceso que trata de la asociación, correlación y combinación de datos e información obtenidos de múltiples o una fuente para realizar estimaciones (posición e identidad) y evaluaciones completas de situaciones y amenazas o impacto y su importancia. El proceso es caracterizado por refinamientos continuos de sus valoraciones y estimaciones junto con la evaluación de la necesidad de fuentes adicionales o modificación del proceso para lograr mejores resultados”.

La fusión de datos ha logrado impactar una gran cantidad de disciplinas y es usada extensamente en áreas como redes de sensores, robótica, procesamiento de señales, de imágenes y videos, diseño de sistemas inteligentes, entre otros (Becerra et al., 2021; Cheng et al., 2013; H. Li et al., 2013; Liang et al., 2019). Por lo tanto, la fusión de datos es un tema de amplio espectro, e incluye una gran variedad de definiciones, las cuales han sido usadas indistintamente, cuyas terminologías y métodos han sido mencionados en patentes (Olabarrieta & Del Ser, 2011; Srivastava et al., 2013) y publicaciones científicas, en áreas de la ingeniería, la medicina, robótica, gestión, defensa, área financiera y de negocios, y muchas otras áreas que utilizan diferentes tipos de datos e información (Becerra et al., 2021). Sin embargo, la fusión de datos continúa siendo una tarea desafiante, ya que los modelos propuestos en ocasiones son poco efectivos o presentan dificultades para mejorar la calidad de la información, debido a que estos no pueden abordar todos los problemas que afectan a los datos, como son la imperfección, la correlación, la inconsistencia y la disparidad (Castanedo, 2013). Adicionalmente, son muy dependientes de la aplicación (Khaleghi et al., 2013) y de la selección de técnicas, lo que incrementa la complejidad del problema (Sidek & Quadri, 2012), siendo esta última contemplada en un gran número de

arquitecturas reportadas en la literatura en el desarrollo de modelos de fusión de datos (Esteban et al., 2005; Nassar et al., 2010).

Con el fin de determinar el desempeño de los modelos de fusión de datos, gran parte de los estudios implementan una etapa de validación, además de la etapa de identificación y estimación, donde establecen sus propias métricas de evaluación enfocadas en aplicaciones específicas como las presentadas en (Clifford et al., 2011; Q. Li et al., 2008). Estas métricas están enfocadas en técnicas convencionales, las cuales funcionan muy bien en ambientes de prueba, pero pueden presentar problemas en la práctica, al ser muy limitadas en términos de evaluación de la calidad de la información y ausentes de generalidad. Los modelos de evaluación de los sistemas de información deben ser lo suficientemente generales para ser implementados en todos los casos en las diferentes áreas de aplicación (I.-G. Todoran et al., 2013).

A pesar de los múltiples estudios reportados, donde las técnicas de fusión de datos han sido ampliamente utilizadas en diferentes áreas, permitiendo expresar aspectos de procesamiento multi-fuente de la información, tanto en los datos como en el procesamiento de la información en diferentes niveles, y permitiendo la combinación coherente de los elementos de procesamiento (algoritmos) con un rendimiento medible y demostrable (Kokar et al., 2004), alcanzar una única arquitectura que proporcione un conocimiento completo de un fenómeno de interés resulta altamente complejo. Lo anterior, debido a la diversidad de los sistemas y a las múltiples observaciones en diferentes tiempos, condiciones y en múltiples experimentos de un objeto o fenómeno, que tienen información de naturaleza muy diversa (Lahat et al., 2014).

La calidad de la información es un término ambiguo, difícil de describir y muy subjetivo. Su evaluación es considerada una tarea de alta complejidad, por lo que (Mendes et al., 2012) consideran que un juicio canónico para cada tarea no es factible y, a pesar de que los sistemas de información ayudan a resolver tareas a los usuarios, estos están orientados a la calidad de los datos, lo que resulta insuficiente, ya que no contemplan adecuadamente los requerimientos del usuario para alcanzar la calidad de información demandada por este (I.-G. Todoran et al., 2013), siendo un aspecto de gran importancia, ya que la interpretación de la calidad es dependiente de quién utiliza dicha información. Así, mientras que un usuario puede considerar la calidad de los datos suficiente para una tarea determinada, puede que no sea suficiente para otra tarea u otro usuario. Además, se debe considerar que diferentes métricas, las cuales son utilizadas en ambientes de prueba, no pueden ser utilizadas en ambientes reales. Por lo anterior, se considera que en términos generales, el principal problema de los sistemas de fusión de datos es la confiabilidad de su evaluación (I.-G. Todoran et al., 2013; I. Todoran et al., 2015), lo cual es considerado un tópico poco explorado y que carece de un estándar que permita una adecuada evaluación de la calidad de dichos sistemas, generando dificultades en los sistemas de fusión actuales en cuanto a su modificación, integración, reutilización y evolución.

Los sistemas de fusión de datos, a pesar de presentar un buen desempeño las etapas de fusión de datos, pueden generar un mal desempeño de los sistemas de información de manera global, afectando negativamente la toma de decisiones. Por lo anterior, no resulta adecuado para este tipo de sistemas aceptar un modelo de evaluación tipo caja negra, como el presentado por (Wang

& Strong, 1996), el cual ha sido ampliamente aceptado. Por lo tanto, se requieren sistemas de evaluación suficientemente completos que intervengan no solo de manera global, sino también de manera local, trazar las causas del mal desempeño y examinar criterios de calidad tanto objetivos como subjetivos (van Laere, 2009). Adicionalmente y de acuerdo a lo presentado por (Hadzagic et al., 2012), es necesario considerar que la evaluación cuantitativa de la calidad de la información y su impacto en el rendimiento de un algoritmo de fusión, facilita la evaluación de la calidad, lo que contribuye a mejorar la valoración de la situación y la toma de decisiones.

En síntesis, se han identificado diferentes aspectos percibidos como problemáticos, los cuales afectan la confiabilidad de la calidad de la información en los sistemas de fusión de datos y su evaluación, y que impactan negativamente el resultado entregado y, por consecuencia, afectan la toma de decisiones: (1) incapacidad de la fusión de datos para abordar todos los problemas que afectan a los datos, (2) conflictos entre las dimensiones de calidad y dependencia al contexto, (3) ausencia de generalidad en las métricas y establecimiento de una medida de calidad estándar, (4) la calidad de la información desde los requerimientos del usuario, (5) trazabilidad de la calidad de la información a través de los niveles de fusión de datos para evaluar su desempeño, y (6) las técnicas en ocasiones son poco efectivas o presentan dificultad para mejorar la calidad de la información.

1.3 HIPÓTESIS

El *framework* de fusión de datos en el marco del modelo JDL y con capacidad de valoración de la calidad de la información, propuesto en el presente trabajo, permite refinar la cadena de procesamiento en los sistemas de información y fusión de datos, mejorando el desempeño y la confiabilidad de la información entregada al usuario para el soporte de decisión.

1.4 JUSTIFICACIÓN

Actualmente la complejidad de los sistemas de información ha aumentado debido a la creciente cantidad de información obtenida desde diferentes ambientes a través de múltiples dispositivos tecnológicos (I.-G. Todoran et al., 2013), razón por la cual la fusión de datos, como herramienta para reducir y extraer información con bajos niveles de incertidumbre, sigue siendo un campo abierto de investigación, el cual es abordado desde diferentes frentes (i.e. fusión de datos, calidad de la información, contexto, estadística, entre otros). Sin embargo, son pocos los estudios que han abordado la fusión de datos y la calidad de la información en conjunto (Rogova & Snidaro, 2018; I. Todoran et al., 2015).

Los sistemas de información son utilizados en múltiples campos como herramientas de apoyo a la toma de decisiones y automatización (Becerra et al., 2021; I. Todoran et al., 2015). El desempeño de estos sistemas depende en gran medida de la calidad de información que procesan, siendo necesario monitorearla de manera permanente para que los sistemas y usuarios puedan realizar ajustes que permitan reducir impactos negativos. Teniendo en cuenta

lo anterior, es posible afirmar que el valor que aporta la información depende de sus características de calidad (Bisdikian et al., 2013), lo cual hace necesario desarrollar sistemas que permitan un manejo más eficiente de la información y que mejoren la calidad de la misma.

Otra de las dificultades que enfrentan los sistemas de información es el creciente volumen de información con variados niveles de relevancia y redundancia, lo cual hace complejo el procesamiento para los sistemas de información (Becerra et al., 2021). Sin embargo, a través de la fusión de datos es posible reducir el volumen de información y aprovechar la redundancia para mejorar la confiabilidad de la información (Castanedo, 2013; Khaleghi et al., 2013). Algunas de las dificultades que presentan los sistemas de información, han sido abordadas por (I. Todoran et al., 2015), a través de una metodología que permite la descomposición de los sistemas en mínimas unidades, las cuales son modeladas a partir de métricas de calidad para entregar al usuario final una trazabilidad de la calidad a través de la cadena de procesamiento de datos existente en un sistema de información. Sin embargo, esta metodología no proporciona un criterio de descomposición, lo cual hace que se pierda generalidad y los resultados de la medición de la calidad no sean utilizados para intervenir o afectar el proceso para disminuir la incertidumbre en la información. Por otro lado, (Rogova, 2016) realiza un análisis teórico de la calidad de la información en los sistemas de fusión de datos, proponiendo un conjunto de criterios de calidad agrupados en 3 categorías: calidad de la fuente (dato), calidad del contenido y calidad de la presentación; logrando cubrir los diferentes niveles del reconocido modelo de fusión de datos JDL e involucrando la calidad de la información del contexto como parte de la propuesta. Mas tarde, (Rogova & Snidaro, 2018) incluyeron en su propuesta la valoración de la calidad de la información como metadatos para mejorar la confiabilidad de las valoraciones de la calidad de la información. A pesar de los esfuerzos realizados para aplicar la calidad de la información sobre los sistemas de fusión de datos, estos estudios aún son muy limitados y presentan algunos vacíos, como es el uso de las medidas de la calidad de la información para afectar los procesamientos en el sistema de información, con el fin de mejorar la calidad de la información a través de una arquitectura general que permita mapear cualquier ambiente. En este sentido, el *framework* en lazo cerrado en el marco del modelo JDL propuesto en este trabajo, aborda dichos vacíos desde tres aspectos: i) el *framework* de evaluación de la calidad de la información responde al problema de la confiabilidad de la evaluación de los sistemas de fusión de datos y la valoración de la calidad de la información reportado en (Dragos & Rein, 2014; I.-G. Todoran et al., 2013; I. Todoran et al., 2015); ii) el *framework* de fusión de datos propuesto presenta adaptabilidad que permite mejorar el procesamiento en los sistemas de información; y iii) la amplitud del espectro del *framework* propuesto para múltiples aplicaciones. En conjunto, estos tres aspectos permiten mejorar el desempeño y la confiabilidad de la información entregada al usuario para el soporte de toma de decisiones.

1.5 OBJETIVOS

Objetivo general

Proponer un *framework* de fusión de datos en lazo cerrado orientado a la calidad de la información en el marco del modelo JDL, para el mejoramiento del desempeño y la confiabilidad de la información entregada al usuario para el soporte de decisión.

Objetivos específicos

- Proponer un conjunto de criterios de calidad de la información, sustentado en experimentos para cada nivel del modelo JDL, a partir de un análisis de diferentes niveles de calidad de la información reportados en la literatura, para determinar su funcionalidad y tipo de datos como parte del nivel de refinamiento.

- Proponer un *framework* de fusión de datos con realimentación, basado en criterios de calidad de la información, que permita la optimización del procesado en cada nivel, para el mejoramiento de su rendimiento de acuerdo con los requerimientos de calidad del usuario.

- Validar el *framework* propuesto en múltiples ambientes de prueba, que permita evaluar su funcionalidad, generalidad y limitaciones a partir de múltiples análisis estadísticos.

2. METODOLOGÍA

En la Figura 2 se muestra la metodología aplicada para el desarrollo de este trabajo, la cual fue ejecutada siguiendo dos fases, una teórica y una experimental.

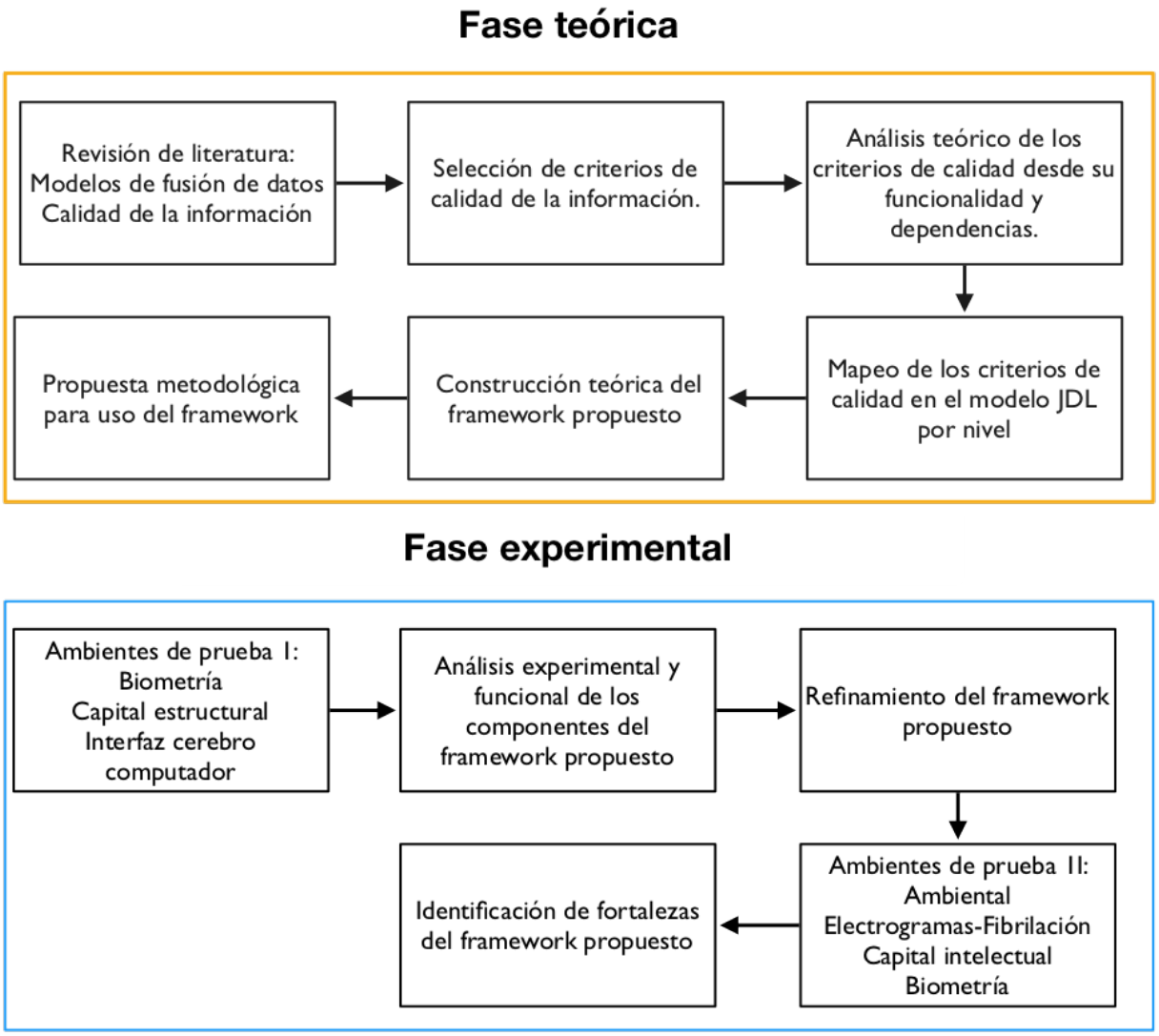

Figura 2. Metodología de la investigación.

2.1 FASE TEÓRICA

En la primera fase se llevó a cabo la revisión de la literatura para la caracterización de los sistemas de fusión de datos en el marco del modelo JDL, a partir de criterios de calidad de la información y los requerimientos de usuario y/o aplicación (ver Artículo 1).

Luego se construyó teóricamente un *framework* inicial de procesamiento de datos realimentado, que combina el modelo JDL y la calidad de la información, considerando su funcionalidad, generalidad, ventajas y limitaciones; con el objetivo de obtener un *framework* con capacidad de adaptación a diferentes casos de estudios o aplicaciones. El *framework* teórico integra las técnicas funcionales de procesamiento de datos, fusión de datos y calidad de la información bajo problemas identificados desde el dato, el usuario y/o máquina. Se mapeo cada nivel del modelo JDL a partir de criterios de calidad de la información ejecutando un análisis de dependencia desde lo teórico. Para cada nivel del *framework* se propuso un modelo de regresión en función de los criterios de calidad de entrada salida, para ser utilizado como una función objetivo *y* optimizar la calidad de la información a través de los diferentes niveles del modelo JDL, aplicando algoritmos de optimización meta-heurísticos. Además, se analizaron diferentes componentes de los algoritmos de procesamiento en el marco del *framework* propuesto con el fin de evaluar sus debilidades y fortalezas en el contexto de múltiples aplicaciones, para determinar su viabilidad de adaptación en el manejo de los diferentes tipos de datos, proponiendo adaptaciones necesarias en la cadena de procesamiento.

2.2 FASE EXPERIMETAL

Con el fin de validar y perfeccionar el *framework* propuesto se dio paso a la fase experimental. Esta fase incluyó la validación del *framework* de procesamiento de datos propuesto, usando diferentes ejemplos que contemplan los niveles de la arquitectura JDL, para identificar las fortalezas y debilidades del *framework*. Primero se aplicó el *framework* en diferentes ambientes de prueba, en el área de la Biometría (ver Artículo 2) y en Valoración del capital humano (ver Artículo 3). Con los resultados preliminares se perfeccionó el *framework* y se procedió a validar en nuevos casos de uso donde intervienen datos de los niveles bajo y alto de acuerdo al modelo JDL, que permitieron medir las fortalezas y debilidades del *framework*. El primer caso de uso corresponde al ambiente de prueba de variables ambientales (ver Artículo 4), el segundo caso de uso corresponde al tratamiento de Electrogramas para la identificación de zonas de ablación en fibrilación auricular (ver Artículo 5). Finalmente, se confrontaron los resultados generados entre ellos y con algunas propuestas de la literatura, a fin de determinar las contribuciones concretas alcanzadas.

A continuación se detalla el *framework* propuesto.

Propuesta: *framework* de procesamiento de datos y calidad de la información para sistemas de información

En la Figura 3 se muestra el *framework* propuesto que a diferencia del modelo JDL original, este cuenta con un bloque de calidad de la información el cual se encarga de valorar la calidad de la información a través de todos los niveles del *framework*. Además cuenta con una sección de identificación del contexto lo cual resulta en un apoyo fundamental para la dinámica y contextualización de cada caso de estudio o aplicación. El *framework* cuenta con una entrada de datos, seguido de una etapa de fusión de bajo nivel la cual corresponde al nivel 0 y 1 del

framework. La etapa de fusión de alto nivel corresponde a los niveles 2 y 3. Los niveles y bloques complementarios son descritos a continuación:

i) Nivel 0 (selección de datos y pre-procesamiento): en este se seleccionan los datos, se filtran eliminando datos defectuosos o atípicos y se aplican técnicas para aproximar datos inexistentes o eliminados basados en históricos de información.

ii) Nivel 1 (asociación, extracción de características, identificación y rastreo): en este nivel se calculan características basadas en medidas obtenidas del fenómeno o proceso y se detectan, identifican y rastrean objetos. En este nivel se utilizan técnicas de transformación de los datos y técnicas de aprendizaje automático que permiten develar o representar la dinámica subyacente del fenómeno comportamental bajo estudio.

iii) Nivel 2 (valoración de la situación): en este se realiza la valoración de la situación en función de los resultados obtenidos en los niveles anteriores, en la calidad de la información y en la información del contexto. Se aplican técnicas de *clustering* para identificar relaciones. Este nivel permite determinar capacidades y vulnerabilidades. También se aplican sistemas de inferencia o razonadores basados en casos para establecer reglas o relaciones que pueden ser obtenidas de expertos o históricos de información, lo cual permite realizar la valoración de la situación y del riesgo en conjunto, teniendo en cuenta la calidad de la información a lo largo de toda la cadena de procesamiento.

iv) Nivel 3 (valoración del riesgo): a partir de los resultados obtenidos en el nivel anterior se puede llevar a cabo la valoración del riesgo o impacto, y definir qué acciones a ejecutar basado predicciones a futuro, oportunidades y vulnerabilidades. Esto puede ser realizado usando los sistemas de inferencia, razonadores basados en casos y predictores.

v) Nivel 4: en este nivel se ejecutan ajustes con el fin de refinar el procesado de los datos, seleccionando fuentes de datos, y ajustando parámetros libres de los algoritmos de procesamiento, detección y predicción basado principalmente en los requerimientos de usuario respecto a la calidad de la información y se basa en resultados entregados por un sistema de optimización que relaciona cuales variables de calidad deben ser afectadas para mejorar la calidad de la información.

vi) Nivel 5 (refinamiento por usuario y HCI): el usuario complementa el refinamiento del proceso manualmente (es opcional) y su intervención puede limitarse a establecer sus requerimientos para poder optimizar el proceso. La interfaz hombre máquina (HCI) corresponde a la presentación de los datos a los usuarios finales.

vii) Calidad de la información: en este bloque se lleva a cabo la asignación de criterios y valoración de la calidad de la información para cada nivel. El módulo de valoración de la calidad de información junto con sus criterios de calidad se explica más abajo.

viii) Administración base de datos: en este módulo se almacenan los históricos de datos que permiten a los modelos dentro del *framework* realizar predicciones adecuadamente y refinar el procesamiento de los datos.

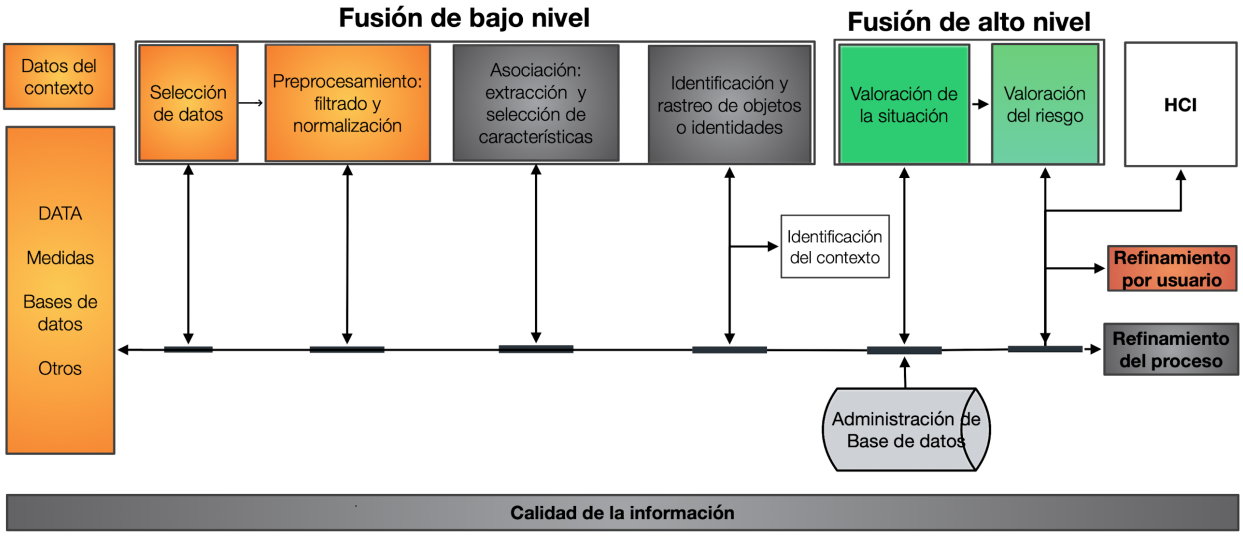

Figura 3. *Framework* propuesto de procesamiento de datos y calidad de la información.

Módulo de valoración de la calidad de la información

La IQ es una variable multidimensional la cual es caracterizada por múltiples criterios y su selección depende de la aplicación y de los requerimientos de usuario (Mendes et al., 2012). Cada criterio debe ir en concierto con las métricas usando información del contexto. Los criterios y medidas no son necesariamente independientes entre ellos, por lo cual se debe realizar un análisis de dependencia para evitar conflictos entre las variables y efectos negativos sobre la valoración de la IQ. Además se debe considerar la existencia de una jerarquía entre los criterios teniendo en cuenta el contexto, aplicación y usuario (Rogova & Bosse, 2010) y su selección debe considerar el costo de su valoración (Haug et al., 2011). En el *framework* propuesto la jerarquía es establecida considerando cada nivel del *framework* y se aplica el uso de la metodología "*Total Data Quality Management-TDQM cycle*" (Wang et al., 2002) la cual se compone de 4 etapas principales (ver Figura 4):

i) Definición: esta identifica las necesidades del usuario y establece los criterios de calidad.

ii) Medición: determina y aplica las métricas de calidad.

iii) Análisis: identifica los problemas e impacto de las medidas de calidad.

iv) Mejoramiento: realiza diferentes procedimientos para mejorar la calidad de la información.

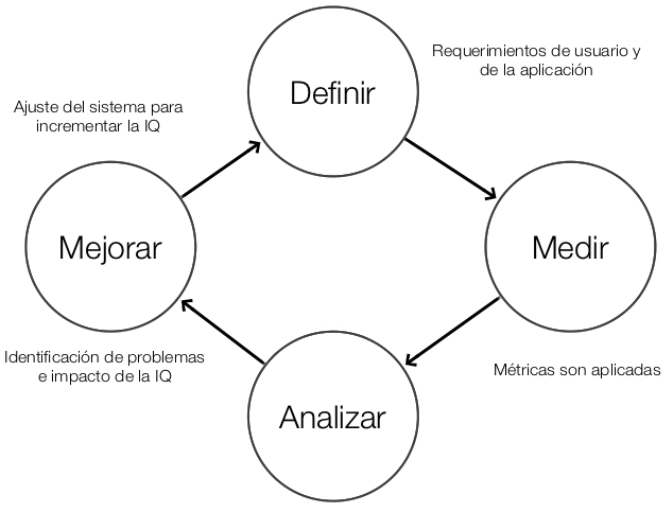

Figura 4. Metodología *Total Data Quality Management-TDQM cycle.*

Esta metodología fue aplicada en un ambiente de fusión de datos para interfaces humano computador (Becerra et al., 2018) para lo cual, en este modelo se sugiere seguir un procedimiento similar. En la Tabla 4 se presentan los criterios de IQ con sus respectivas métricas propuestas para cada uno de los niveles del modelo. Estas fueron seleccionadas teniendo en cuenta la funcionalidad de cada uno de los niveles del *framework* y refinada a partir de los diferentes casos de estudio llevados a cabo en este trabajo.

Tabla 4. Criterios y métricas de calidad de la información

Nivel	Criterio	Métrica
L0, L1	Precisión	Atributos del dato para una adecuada discriminación en el contexto.
L0, L1	Exactitud	Aproximación al valor real
L0	Cantidad de datos	Resolución de la medida
L0	Completitud	% de muestras, campos, registros respecto al total esperado.
L1	Consistencia	Distancia de los datos obtenidos desde diferentes fuentes (dependencias) y que se encuentre dentro del rango.
L0	Verificabilidad	Es el nivel de la posibilidad de testear la correctitud de los datos.
L0, L1	Reputación	Reputación de la fuente desde donde fueron obtenidos los datos. Funcionalidad y robustez de características.
L0, L1	Volatibilidad	Rango de tiempo durante el cual la información tiene validez
L0, L1	Relevancia	Importancia de las medidas o características para el proceso de identificación o tracking.
L0, L1	Consistencia	Relación coherente entre variables
L1	Robustez	Nivel de relevancia de las características respecto al nivel de ruido. Ponderación obtenida por el selector *Relief-F.* $\sum_{ruido=1}^{n} Wn$
L1	Confiabilidad	Precisión de cada predictor a partir de su análisis

L2-L3	Objetividad	Confianza de la valoración del riesgo.
L2	Reputación	Ponderación dada a las reglas en el FIS.
L2-L3	Calidad Local	*QL*= $\sum_1^n Wn\ Qn$ donde *Qn* corresponde al vector de critérios de calidad
L2–L3	Calidad Global	*QG*= $\sum_0^3 Wi\ QLi$
L5	Interpretabilidad	Facilidad de interpretación por parte del usuario final
L5	Autoridad	Nivel confianza de la información entregada por los expertos.
L0-L5	Eficiencia	Costo de recurso respecto al cumplimiento de la tarea.

La estrategia para la valoración de la calidad de la información a través de los niveles del *framework* propuesto junto con el proceso de optimización es explicada a continuación:

Valoración y trazabilidad de la calidad de la información

La valoración de la calidad de la información es ejecutada estimando los criterios presentados en la Tabla 4 a partir de sus respectivas métricas representadas por el vector $\boldsymbol{q}$ (este contiene las métricas de calidad para cada dato o variable) y es calculado por cada nivel. Las $\boldsymbol{IQ_{Lx}}$ (calidades locales) para el nivel 0 ($\boldsymbol{IQ_{L0}}$), el nivel 1 ($\boldsymbol{IQ_{L1}}$), el nivel 3 ($\boldsymbol{IQ_{L2}}$) y el nivel 3 ($\boldsymbol{IQ_{L2}}$) es calculada usando la ecuación (1) pero ajustando los parámetros de ponderación ($\boldsymbol{w_{xy}}$) para cada nivel. El nivel 2 y 3 aplican la ecuación (2) pero con sus respectivos parámetros de ponderación y métricas de calidad $\boldsymbol{q}$. El número de métricas $\boldsymbol{s}$ puede diferir entre cada nivel.

$$IQ_{Lx}=w_{G1}IQ_{Lx-G1}+w_{G2}IQ_{Lx-G2}+\cdots+w_{Gn}IQ_{Lx-Gn} \qquad (1)$$

Donde w_{Gn} corresponde a la ponderación dada a cada criterio. Los *IQ* son definidos como:

$$\begin{aligned} IQ_{Lx-G1} &= w_{11x}q_{G11x,cr}+w_{12x}q_{G12x,cr}+\cdots+w_{1mx}q_{G1xm,cr}\,, \\ IQ_{Lx-G2} &= w_{21x}q_{G21x,cr}+w_{22x}q_{G22x,cr}+\cdots+w_{2px}q_{G2px,cr}\,, \\ IQ_{Lx-Gn} &= w_{n1x}q_{Gn1x,cr}+w_{n2x}q_{Gn2x,cr}+\cdots+w_{nrx}q_{Gnrx,cr}, \end{aligned} \qquad (2)$$

donde *m*, *p* y *r* son el número total de métricas establecidas para el grupo 1, 2 y *n,* respectivamente. *Cr* corresponde al criterio evaluado por el conjunto de métricas.

La trazabilidad de la *IQ* es llevada a cabo a partir de la generación de funciones que relacionan las entradas y la valoración de su calidad respecto a las salidas y su calidad. Lo anterior se realiza en todos los niveles como se ilustra en la Figura 5. En cada nivel se obtiene una función de regresión en función de los criterios de calidad. Estas funciones permiten realizar una predicción de la calidad de la información, mantener una trazabilidad y poder realizar procesos de optimización para mejorar las IQ en las salidas de cada uno de los niveles del *framework* propuesto. La inclusión de la información como la tupla (I_{nr}, q_{nr}) o solo la valoración de la calidad de la información ($q_{n,r}$) depende del caso de estudio.

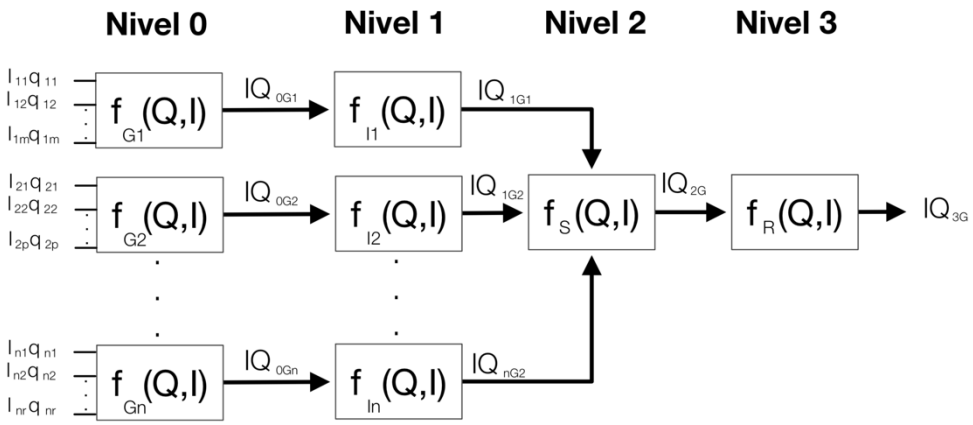

Figura 5. Optimización de la calidad - Trazabilidad de la calidad de la información.

Metodología para aplicación del *framework* propuesto

En la Figura 6 se muestra el conjunto de pasos a seguir para la aplicación del *framework* propuesto en cualquier ambiente de prueba. Primero se debe analizar el ambiente de estudio y realizar agrupaciones ya sea por funciones o procesos mapeando los procesos en cada nivel del *framework* propuesto y así proceder a ejecutar los siguientes pasos:

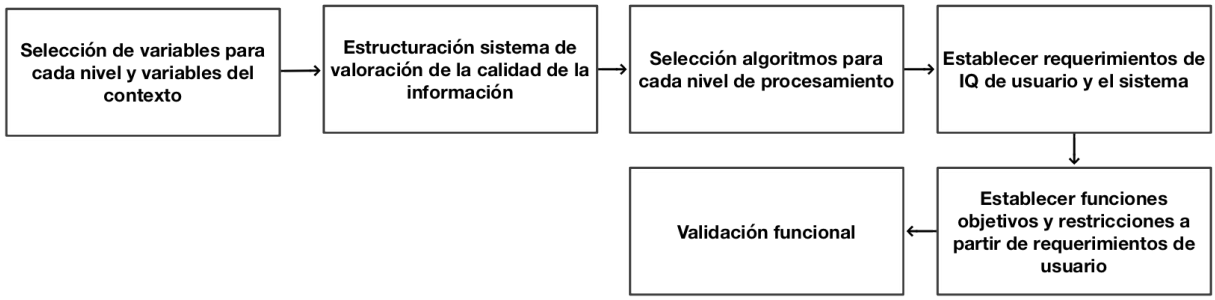

Figura 6. Descripción metodológica de uso del *framework* propuesto.

i) Selección de variables para cada nivel y variables de contexto: se debe de determinar la disponibilidad de datos e información y establecer variables de interés basado en conocimiento a priori. Basado, en la funcionalidad de cada nivel se debe de determinar variables de interés a identificar en el nivel 1, posibles relaciones para valorar la situación en el nivel 2 y se deben de determinar riesgos e impacto posibles en el nivel 3. Estos dos últimos niveles requieren el acompañamiento de expertos en el área de estudio.

ii) Estructuración del sistema de valoración de la calidad de la información: Esta es llevada a cabo siguiendo la metodología descrita en la Figura 7. Primero se deben de caracterizar los datos por medio de funciones de probabilidad o de posibilidad. Luego deben ser caracterizado cada nivel en función de las variables de entrada y de los resultados entregados por los niveles anteriores al nivel que se está caracterizando. Luego se seleccionan un conjunto de criterios de calidad de la información junto con sus métricas

(estas tienen alta dependencia con el caso de estudio). Luego se realiza un análisis de imperfección de los datos y se adiciona ruido a los datos generando diferentes bases de datos con la valoración respectiva de la calidad de la información y a partir de estos resultados se generan funciones de regresión para predecir la IQ para cada nivel del *framework*. Se establecen criterios relevantes para el caso de estudio usando análisis de correlación y selección de características usando algoritmos tipo filtro o *wrapping.*

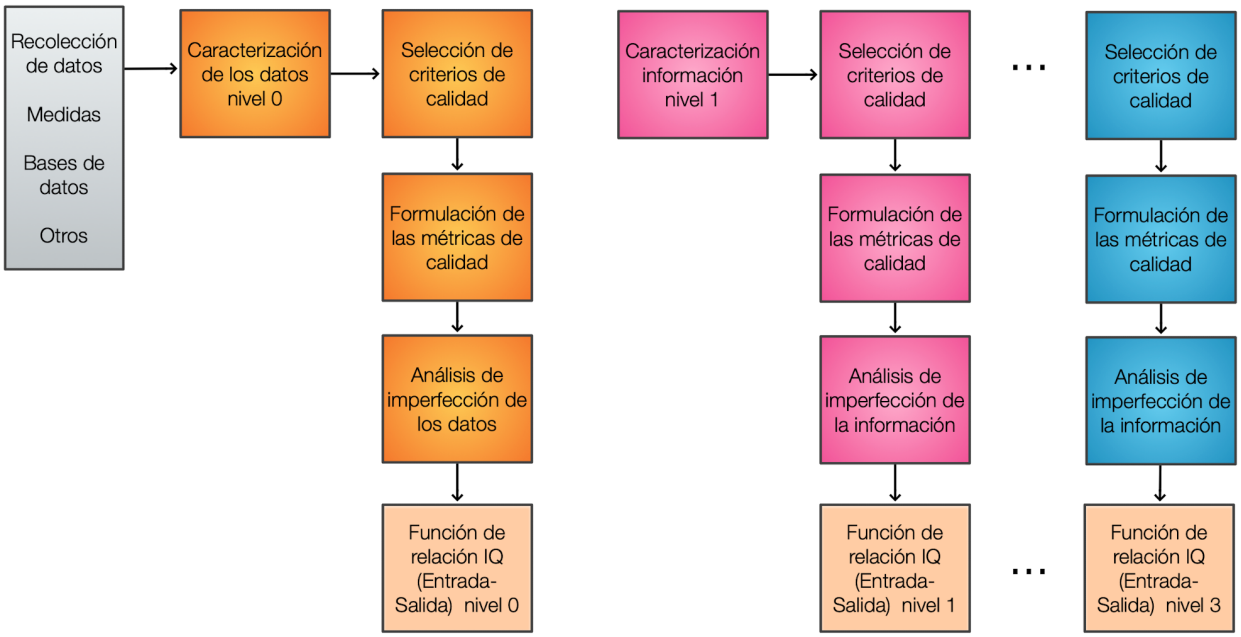

Figura 7. Metodología para construcción de la etapa de valoración de la calidad de la información

iii) Selección de algoritmos para cada nivel de procesamiento: estos son seleccionados teniendo en cuenta el ambiente de prueba. Las grandes diferencias en este proceso se evidencian en el nivel 1 para los diferentes ambientes de prueba. Para los niveles 2 y 3 se recomienda el uso de sistemas basados en reglas o razonadores en basado en casos, para valorar la situación y el riesgo/impacto a partir de la información obtenida del contexto, la información de los niveles previos y la IQ.

iv) Establecer requerimientos de usuario y del sistema en función de la calidad de la información.

v) Establecer funciones objetivo y restricciones en función de los requerimientos de usuario. El algoritmo de optimización dependerá del comportamiento de la calidad de la información. Para obtener una generalidad se recomienda el uso de un algoritmo de optimización meta-heurístico pero en algunas ocasiones es suficiente el uso de un optimizador lineal.

vi) Validación funcional.

En los casos de estudio se ejemplifica el funcionamiento del *framework* propuesto junto con su metodología de aplicación.

3. RESULTADOS

3.1 FASE TEÓRICA

La revisión de literatura realizada en este trabajo es presentada en el siguiente artículo:

Information Quality Assessment Oriented to Data Fusion Systems (Artículo 1): Este artículo da cuenta de la revisión del estado del arte de esta tesis y soporta los conceptos, hipótesis y teorías utilizadas. Un análisis comparativo de los diferentes modelos de fusión de datos es llevado a cabo para establecer la generalidad, ventajas y desventajas del modelo JDL utilizado en esta tesis. Adicionalmente, se analizan múltiples metodologías para la valoración de la calidad de la información y se estudian los diferentes criterios de calidad que han sido propuesto lo cual permitió seleccionar los criterios de calidad para la construcción de nuestro *framework* propuesto desde lo teórico. Finalmente, se estudiaron las metodologías propuestas hasta la fecha que combinan tanto fusión de datos como calidad de la información con el fin de identificar ventajas, desventajas, limitaciones y desafíos done algunos de ellos son abordados en el desarrollo de este trabajo.

3.2 FASE EXPERIMENTAL: APLICACIONES DEL FRAMEWORK PROPUESTO

A continuación, se describen los artículos asociados a la fase I y II. Los artículos correspondientes a la fase permitieron realizar un estudio teórico del modelo de fusión JDL, de la calidad de la información y análisis exploratorios de los ambientes de prueba realizando mapeos del modelo JDL sobre el ambiente de prueba, además de construir *frameworks* pilotos que permitieron encontrar fortalezas y debilidades para refinar la propuesta y finalmente validar este primer *framework* en los ambientes de prueba de la fase 2.

Ambientes de prueba I

Modelo JDL y calidad de la información para identificación biométrica a partir de señales multimodales - estudio exploratorio (Artículo 2): Este artículo corresponde a la aplicación del f*ramework* propuesto en un caso de identificación biométrica usando datos no estructurados i.e. señales fisiológicas multimodales. El trabajo permitió definir criterios de calidad de la información con sus respectivas métricas para cada nivel, aplicado a datos no estructurados. También, permitió analizar los efectos de la calidad de la información y el mapeo del *framework* en el ambiente de identificación biométrica, el cual incluye procesamiento de los datos en bruto, detección de patrones, valoración de la situación y del riesgo o impacto. Los resultados demostraron la funcionalidad del modelo propuesto y su potencialidad respecto a otros modelos de identificación tradicionales considerando la valoración del riesgo. También se identificaron fortalezas y debilidades de algunos algoritmos de predicción frente a variaciones en la calidad de la información y atributos con mayor tolerancia al ruido.

Modelo de capital estructural para Universidades basado en el modelo de fusión de datos JDL y la calidad de la información (Artículo 3): Este artículo corresponde a la aplicación del framework propuesto en un caso de capital estructural. Este artículo al igual que el anterior permitió poner a prueba el *framework* propuesto. Este trabajo se enfocó a la valoración del capital estructural en instituciones de educación superior. Este ambiente maneja datos estructurados lo cual permitió identificar fortalezas y limitaciones del *framework* propuesto al manejar este tipo de datos. Se identificaron criterios y métricas de calidad para cada nivel del *framework* y permitió realizar ajustes en la metodología de la aplicación del framework. El modelo construido a partir de este *framework* no solo permitió la valoración del capital estructural, sino también apoya la toma de decisiones basada en la calidad de la información y su impacto. También se propusieron criterios de valoración del capital estructural por nivel de particular uso de este ambiente. Se identificaron desafíos como la consideración de otros criterios de calidad de la información y valoraciones de tipo longitudinal.

Ambientes de prueba II

Information fusion and information quality assessment for environmental forecasting (Artículo 4): Este artículo corresponde a la aplicación del *framework* propuesto en un caso de pronóstico ambiental. Un análisis del efecto del ruido sobre datos ambientales obtenidos desde múltiples fuentes fue ejecutado, para determinar los efectos sobre la calidad de la información sobre los diferentes niveles del *framework* propuesto. La funcionalidad del *framework* fue validada, la relevancia de los criterios de calidad fue establecida junto con sus métricas.

Data fusion and information quality model for atrial fibrillation analysis from electrograms (Artículo 5): Este artículo corresponde a la aplicación del *framework* propuesto en un caso de estudio de fibrilación auricular a partir de electrogramas. Un estudio de calidad de la información es aplicado sobre electrogramas (EGM) y mapas generados a partir de características obtenidas a partir de los EGM con el fin de ver los efectos sobre la trazabilidad de la calidad de la información, tolerancia al ruido y efecto sobre la localización del punto espacial que origina un rotor que devela la fibrilación auricular. Además se aplica fusión de datos a nivel de la señal y nivel de los mapas considerando los resultados de la calidad de la información para mejorar el resultado de la localización del tip del rotor y se realiza el proceso de optimización basado en las medidas de calidad en función de requerimientos de usuario. También se propone la valoración de la situación y del riesgo. Los resultados permitieron demostrar la funcionalidad del *framework* y establecer los criterios de calidad (con sus respectivas métricas) relevantes en este ambiente de prueba.

En la tabla 5 se resumen las principales contribuciones de esta tesis:

Tabla 5. Contribuciones doctorales al tema propuesto

Contribuciones doctorales al tema propuesto	Aportes del framework propuesto al problema
Framework de propósito general en donde se evalúa la calidad de la información y a partir de ella se optimiza el procesamiento de los datos, con el fin de mejorar la calidad de la información demandada por el usuario final. Sistema de procesamiento de información que integra la fusión de datos, máquinas de aprendizaje, optimización, calidad de la información.	-Se identificaron aspectos claves en la aplicación del *framework* en diferentes ambientes de prueba, lo cual permitió obtener un *framework* robusto y de alta adaptabilidad. -Validación de funcionalidad en múltiples ambientes de prueba, demostrando la generalidad y capacidad del *framework* propuesto. Y capacidad de refinar la cadena de procesamiento.
Fase teórica se centró en el estudio de las metodologías, modelos y arquitecturas para la valoración de la calidad de la información y su aplicación en la fusión de datos a través de una revisión de literatura exhaustiva.	-Este trabajo se puede considerar una evolución de la metodología propuesta por Todoran generalizada en un *framework* de fusión de datos en el marco del modelo JDL con un significativo valor agregado en la optimización y valoración de la calidad de la información por niveles funcionales. -Identificación de la necesidad de proponer una metodología de valoración de calidad de la información orientada al usuario final en los sistemas de información con trazabilidad y simplicidad en su aplicación o uso.
Framework basado en el modelo JDL considerando su funcionalidad y cubrimiento de todos los niveles de procesamiento que intervienen en los sistemas de información.	- Metodología de uso del framework. - Set de criterios y métricas de calidad propuestos tanto generales como específicos a algunas aplicaciones. - Modelado de la calidad de la información por niveles. - Múltiples conceptos y desarrollos son reutilizados en este trabajo, pero analizados de forma rigurosa en cada caso de estudio. - Los 6 niveles del modelo JDL fueron caracterizados y modelados usando técnicas de inteligencia computacional, Machine Learning y lógica difusa, a partir de un conjunto de criterios y métricas de calidad propuestas para cada nivel basado en su funcionalidad. - Proceso de optimización multi-objetivo a través de todos los niveles del framework considerando evaluaciones locales y global de la calidad de la información. El framework propuesto fue validado en 4 ambientes de soporte de decisión.

A continuación se presentan los artículos que componen la presente Tesis Doctoral.