Framework para la gestión de calidad de datos

Diego Oliveira Rizzo

Nahir Toledo Olivera

Tutores

MSc. Ing. Flavia Serra

Dra. Ing. Adriana Marotta

Instituto de Computación - Facultad de Ingeniería.

Universidad de la República

Montevideo, Uruguay

Julio 2022

Agradecimientos

Queremos extender nuestro agradecimiento a nuestras familias y amigos por el apoyo incondicional durante estos años de formación. Cada una de estas personas nos han motivado y acompañado a lo largo de este tiempo.
A la profesora Andrea Delgado por su tiempo brindado para guiarnos en el uso de la herramienta Eclipse Process Framework Composer.
A nuestras tutoras, Flavia Serra y Adriana Marotta, por la paciencia, enseñanzas y dedicación en este proyecto.

"Even though quality cannot be defined, you know what it is"
- Robert M. Pirsig

Resumen

La gestión de calidad de datos en las organizaciones implica un conjunto muy grande de tareas, como "*data profiling*", definición de un modelo de calidad, implementación y ejecución de mediciones de calidad, evaluación de calidad teniendo en cuenta requerimientos, monitoreo de la calidad, entre otros. Además, estas tareas son realizadas por personas con distintos roles dentro de una organización. No existe una metodología estándar o globalmente aceptada por la academia o la industria para la organización y realización de estas tareas. En la literatura del área de Calidad de Datos existen algunas propuestas de metodologías o estrategias, así como de frameworks concretos para aplicar ciertas metodologías.

En este proyecto se realiza el estudio de las principales metodologías y frameworks propuestos para la gestión de calidad de datos, y en base a éstos se propone un framework general, que incluya lo más relevante de los frameworks existentes. Este framework se define utilizando la herramienta *Eclipse Process Framework Composer*, y el resultado sirve como guía y asiste en la aplicación del framework. A su vez, este proyecto tiene como referencia una tesis de doctorado en curso que trabaja en el tema de calidad de datos basada en contextos.

Palabras Clave: Calidad de datos, Gestión de calidad de datos, Framework de gestión de calidad de datos.

Índice general

Capítulo 1

Introducción

En este capítulo se presenta una visión inicial del trabajo realizado en este proyecto. En primer lugar, se describe la problemática de una mala calidad de datos y como esto influye en las organizaciones. Luego, se detallan la motivación y objetivos del proyecto, haciendo énfasis en la necesidad de contar con un framework de uso genérico para la gestión de calidad de datos. Por último, se presenta la organización general del documento.

1.1. Problemática y contexto del trabajo

En los últimos años, el continuo avance de las tecnologías de la información ha llevado a que las organizaciones tengan en su poder grandes cantidades de datos. Consecuentemente, los datos se convirtieron en uno de los activos más importantes y valiosos de las organizaciones. Esto se debe a que el correcto análisis de estos datos puede lograr una ventaja competitiva en su negocio. Los datos actúan como guía para los procesos de una organización, por ende, la calidad de estos es de suma importancia [1].

Lograr una buena calidad en los datos no es tarea sencilla. Esto se debe a que no implica únicamente trabajar con datos que no contengan errores, sino que involucra características propias de los datos y que dependen del dominio en donde se utilicen. La mala calidad en los datos trae serias consecuencias en la eficiencia y eficacia de las organizaciones [2]. Por esto, es necesario gestionar los datos para poder evaluar la calidad de estos, mejorarlos y así, agregar valor a los procesos de las organizaciones.

La gestión de calidad de datos implica la planificación de actividades como medir, analizar y controlar distintos aspectos de la calidad de datos que involucra un conjunto de métodos, técni-

cas y herramientas [2]. En la literatura del área de calidad de datos existen muchas propuestas de metodologías, así como también de frameworks que tienen como objetivo gestionar la calidad de datos para determinados dominios [3][4][5][6][7]. Sin embargo, uno de los principales problemas al optar por un framework de gestión de calidad de datos es que no existe un estándar o un framework, de uso genérico para la gestión de calidad de datos, globalmente aceptado por la academia o la industria. Contar con un estándar o framework de uso genérico para las organizaciones tiene distintas ventajas, como lo son la simplificación de los procedimientos de trabajo, la facilitación de la mejora continua y el aseguramiento de la calidad de los productos.

En este proyecto se estudiarán algunos de los frameworks y metodologías disponibles enfocados en la calidad de los datos, y a partir de allí, se definirá un framework para la gestión de calidad de datos de uso genérico. Este trabajo toma como referencia la investigación hecha por Flavia Serra en su trabajo de doctorado en curso [8], a partir de la cual se obtuvo el material bibliográfico que fue punto de partida para este proyecto.

1.2. Objetivos

A continuación, se presentan los objetivos planteados en este proyecto:

- Estudiar los conceptos de calidad de datos.
- Analizar un grupo de metodologías y frameworks seleccionados, ya existentes, aplicados en la gestión de calidad de datos.
- Proponer un framework genérico para la gestión de calidad de datos, tomando como base la bibliografía revisada.
- Utilizar la herramienta *Eclipse Process Framework Composer* [9] para generar documentación que sirva como guía para la aplicación del framework.

1.3. Organización del documento

El presente documento se organiza en siete capítulos. El Capítulo 2 presenta los conceptos y características fundamentales para abordar el área de calidad de datos. El Capítulo 3 realiza un relevamiento y posterior análisis de algunos frameworks y metodologías enfocadas en el área de calidad de datos. Esta investigación es la que sustenta la creación del framework. En el Capítulo 4 se da a conocer la solución propuesta presentando el framework para la gestión de calidad de datos. El Capítulo 5 presenta la implementación del framework definido en el capítulo anterior utilizando la herramienta *Eclipse Process Framework Composer*. El Capítulo 6 presenta un ejemplo de la aplicación del proceso de mejora de calidad de datos propuesto, utilizando un caso de estudio real. Por último, en el Capítulo 7 se presentan las conclusiones obtenidas al finalizar el trabajo, así como también sus aportes y posibles trabajos a futuro.

Capítulo 2

Fundamentos teóricos

El presente capítulo describe los conceptos del área de calidad de datos que guiaron la etapa de investigación de este proyecto. Es importante el entendimiento de estos conceptos para lograr una correcta lectura e interpretación de los siguientes capítulos, así como también, para que el lector pueda comprender las decisiones tomadas durante el desarrollo del proyecto. En la Sección 2.1 se definen los términos dato, información y calidad de datos. Estos términos son clave y son la base para las siguientes secciones. La Sección 2.2 trata sobre modelos de calidad de datos y sus principales características. Por último, la Sección 2.3 introduce qué es la gestión de calidad de datos, el rol de los frameworks y metodologías y algunas herramientas involucradas para dicha gestión.

2.1. Calidad de datos

En los últimos años los datos se han convertido en uno de los activos más importantes para las organizaciones. Su importancia se basa en que son utilizados para realizar análisis que apoyan la toma de decisiones y el diseño de estrategias de negocio e implementación de cambios de forma segura y acertada [10]. Los datos actúan como guía de los procesos de cualquier tipo [11], por lo tanto, es imprescindible contar con datos de buena calidad. A pesar de su importancia, no existe un acuerdo sobre el significado del término "dato" y, comúnmente se utiliza como sinónimo de "información"[12]. A continuación, se presentan dichos conceptos y sus principales características según varios autores.

Tanto científicos como filósofos han utilizado el término "dato" a lo largo de los años [13]. Muchos especialistas en sistemas de información definen el término "dato" como la materia prima a partir de la cual se desarrolla la información. En otras palabras, los datos se utilizan como la entrada de distintos procesos cuya salida depurada es un producto de información [12][14].

De acuerdo con el curso "Calidad e integración de datos" brindado por el Centro de Posgrados y Actualización Profesional en Informática (CPAP) [2], un dato es la representación de un objeto del mundo real elaborado por procesos de software que tiene un formato almacenable, recuperable y que puede ser comunicado a través de una red. Por otro lado, la *International Standard Organization* (ISO) [15], define a los datos como la representación de la información de una manera adecuada para la comunicación, interpretación o procesamiento.

Existen distintos criterios para clasificar a los datos. Entre estos, se destaca su representación, su visión como producto y su complejidad. Batini *et al.* distingue tres tipos de datos según su representación, que van desde datos no estructurados a tipos de datos con una estructura determinada [16].

- **Datos estructurados:** Son agregaciones o generalizaciones de los elementos descritos por atributos definidos dentro de un dominio. Las tablas relacionales constituyen uno de los tipos más comunes de estructuras que contienen datos estructurados.
- **Datos no estructurados:** Son una secuencia genérica de símbolos, típicamente codificados en lenguaje natural que no están asociados a una estructura o tipo de dominio alguno. Un ejemplo típico de datos no estructurados son los cuestionarios con textos libres que responden a preguntas abiertas, es decir, una cadena en lenguaje natural.
- **Datos semiestructurados:** Son datos que tienen una estructura con cierto grado de flexibilidad. Usualmente se representan mediante el lenguaje XML[1].

Las diferencias en la representación de los datos se ven reflejadas en los métodos y técnicas que las organizaciones utilizan para evaluar y mejorar la calidad de los mismos.

Por otro lado, la ISO define el término "información" como el conocimiento sobre conceptos, objetos, hechos, eventos, procesos o ideas que, en un determinado contexto, tienen un significado particular [15]. Aunque la información tiene una representación que la hace comunicable, es su interpretación lo que la hace relevante. A su vez, la información también se define como los datos procesados. Esta noción se corresponde con la comprensión de que la información es un producto que se fabrica a partir de datos, que son considerados como materia prima [17]. La información debe gestionarse de forma adecuada a lo largo de su ciclo de vida de forma tal de obtener el máximo provecho.

[1]XML es un lenguaje de marcado similar a HTML. A diferencia de otros lenguajes de marcado, XML no está predefinido, por lo que debe definir sus propias etiquetas.

En resumen, tal como lo expresa el autor en [1], "puede asumirse que el término dato es usualmente utilizado para referirse a la información en etapas tempranas del tratamiento de un conjunto de datos con un fin específico, mientras que se utiliza el término información para cualquier etapa, por lo que también incluye a los datos. En otras palabras, toda información es un dato en sí mismo, pero no todos los datos constituyen información".

La **calidad de datos** es un concepto muy amplio que está sujeto a diferentes definiciones e interpretaciones dependiendo de los procesos de diseño y producción involucrados en la generación de datos [18]. Es un factor clave ya que representa a la información de una manera formal y adecuada para la comunicación, interpretación o procesamiento [19]. Intuitivamente, una buena calidad de datos consiste en tener datos adecuados y sin errores, en el momento y lugar apropiado donde se desee realizar una tarea específica [1]. En la literatura se utiliza el concepto *fitness for use* (aptitud para el uso) que implica que un conjunto de datos es de buena calidad si son aptos para el uso de los consumidores [20]. Existen otros enfoques donde se indica que, un conjunto de datos es de buena calidad si se ajusta a cierta especificación realizada previamente a su uso. A modo de ejemplo, "en un requerimiento podría ser que para cierto atributo los datos deben cumplir con una determinada regla sintáctica, si los datos disponibles se ajustan a dicha especificación se dice que son de buena calidad"[1]. De lo anterior, se desprende que la calidad de datos es un concepto subjetivo y dependiente del contexto que, a grandes rasgos (y ante la visión del consumidor) deben cumplir con los siguientes requisitos:

- Ser relevantes para su uso.
- Ser correctos y no tener inconsistencias.
- Estar lo más actualizados posible.
- Ser visualizados adecuadamente.
- Ser accedidos fácilmente.

Por otro lado, la mala calidad de los datos impacta negativamente en las organizaciones, ya que deriva en la toma de decisiones erróneas, además de generar insatisfacción por parte de clientes y empleados [1]. Los datos incorrectos se pueden generar de muchas formas, ya sea por errores en la entrada de datos, datos inconsistentes con la realidad o inconsistentes entre sí. También es común encontrar datos desactualizados o información incompleta que impide determinar si los datos son válidos. Poseer datos de fuentes poco confiables, así como también, contar con distintos valores para la misma información es otro de los problemas de interés en calidad de datos [2][21].

Los problemas de calidad de datos se generan en algún punto del ciclo de vida de la información, es decir, durante la producción, el procesamiento, almacenamiento y utilización. Para cada una de estas etapas existen posibles causas que generan los problemas en la calidad de los datos. A continuación, se enumeran los más relevantes [2]:

Producción de los datos:

- Recolección de datos mediante ingreso humano.
- Diferentes fuentes con distintas representaciones del mismo objeto de la realidad.
- Desactualización de los datos.

Procesamiento:

- Unión de datos provenientes de varias fuentes.
- Transformaciones a otras estructuras y formatos.
- Cálculos con datos de entrada, como resúmenes y cálculos de indicadores.

Almacenamiento:

- Formatos diferentes en los datos.
- Ausencia de formatos definidos en los datos.
- Bases de datos mal diseñadas.

Utilización:

- Cambios en los requerimientos de calidad.
- Problemas de seguridad y acceso.
- Capacidad de análisis y procesamiento insuficiente.
- Uso equivocado de los datos, por mala interpretación o aplicación fuera de contexto.

La mala calidad de los datos trae graves consecuencias que influyen en la eficiencia y efectividad de las organizaciones. En muchos casos tiene como consecuencias pérdidas irrecuperables. Existen consecuencias que se pueden observar de forma directa, mientras que otras se manifiestan a lo largo del tiempo. Entre las consecuencias directas se destacan las entregas a los clientes de forma tardía o equivocada, problemas en la implantación de nuevos sistemas de información

que provienen de varias fuentes de datos o información duplicada, entre otros. Para las consecuencias de mediano/largo plazo se encuentra la insatisfacción de los clientes, así como también de las organizaciones, ya que incentiva la desconfianza. Por otro lado, la mala calidad en los datos provoca costos altos e innecesarios, además de impactar en la toma de decisiones [11].

En resumen, la buena calidad en los datos asegura un mejor análisis y la toma de decisiones de negocio de forma segura y eficiente. Sin embargo, garantizar una buena calidad no es tarea sencilla. Es de suma importancia identificar las causas de los problemas de calidad de datos para poder mitigar posibles pérdidas y minimizar las consecuencias mencionadas anteriormente.

2.2. Modelo de calidad de datos

Los modelos son utilizados en distintas disciplinas para representar conceptos, relaciones y procesos, entre otros. En esta sección se describen las principales características de los modelos utilizados para representar la calidad de los datos. Diversos autores coinciden en que la calidad de datos es un concepto multidimensional [22], por lo que, los datos tienen diversas características de calidad, cuya relevancia depende del contexto donde se apliquen.

Un modelo de calidad de datos define el conjunto de datos objetivo, sus características relevantes y cómo medirlas. Para construir un modelo de calidad de datos es necesario definir algunos conceptos de calidad. Dichos conceptos componen la jerarquía presentada en la **Figura 2.1**.

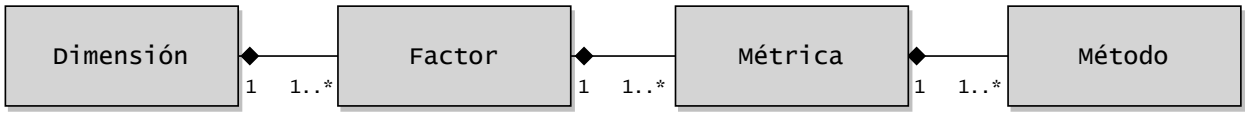


Figura 2.1: Conceptos de calidad de datos. Extraído de [2].

Una **dimensión de calidad** hace referencia a un conjunto de atributos que representan una faceta específica de la calidad de los datos [20], como por ejemplo la exactitud. Las dimensiones definen a alto nivel las características de calidad que deben satisfacer los datos. Según Batini *et al.* [16], las dimensiones de calidad se pueden clasificar en cuatro categorías:

- Intrínseca: Esta categoría refiere a la calidad que los datos tienen por sí mismos. Un ejemplo es la dimensión exactitud.
- Contextual: Hace referencia al contexto en el que son usados los datos. Un ejemplo es la dimensión completitud.
- Representacional: Refiere a aspectos relacionados a la calidad de la representación de los datos. Un ejemplo es la dimensión interpretabilidad.
- De accesibilidad: Refiere a las distintas formas de poder acceder a la información. Un ejemplo es la dimensión accesibilidad.

En el apéndice A se presenta en detalle una descripción de las dimensiones más relevantes según la bibliografía revisada.

Un **factor** representa un aspecto particular dentro de una dimensión de calidad (por ejemplo, exactitud semántica o exactitud sintáctica para la dimensión exactitud). Una **métrica** define de qué forma se debe medir un factor de calidad. Para definir una métrica se debe especificar la unidad de medición (tiempo en ms, volumen en GB, etc) y su granularidad (celda, tuplas, tabla, etc). Por último, un **método de medición** es el proceso que implementa dicha métrica.

En resumen, un modelo de calidad de datos define las dimensiones, factores, métricas y métodos de medición que se implementarán para medir la calidad de un conjunto de datos específico.

2.3. Gestión de la calidad de datos

La gestión de la calidad de datos es la tarea de medir, analizar, mejorar y controlar diferentes aspectos de la calidad de datos que son de interés para un escenario específico [2]. Su principal objetivo es mejorar la calidad de los datos y, por ende, agregar valor a los procesos comerciales [23][24]. Sin embargo, no se busca obtener una calidad de datos perfecta, sino lograr datos de buena calidad según el contexto donde se utilicen [1].

La gestión de calidad de datos involucra un conjunto de tareas, métodos, técnicas y herramientas que permiten administrar la calidad de los datos de un sistema de información [2][25]. Estos elementos se presentan en la **Figura 2.2**.

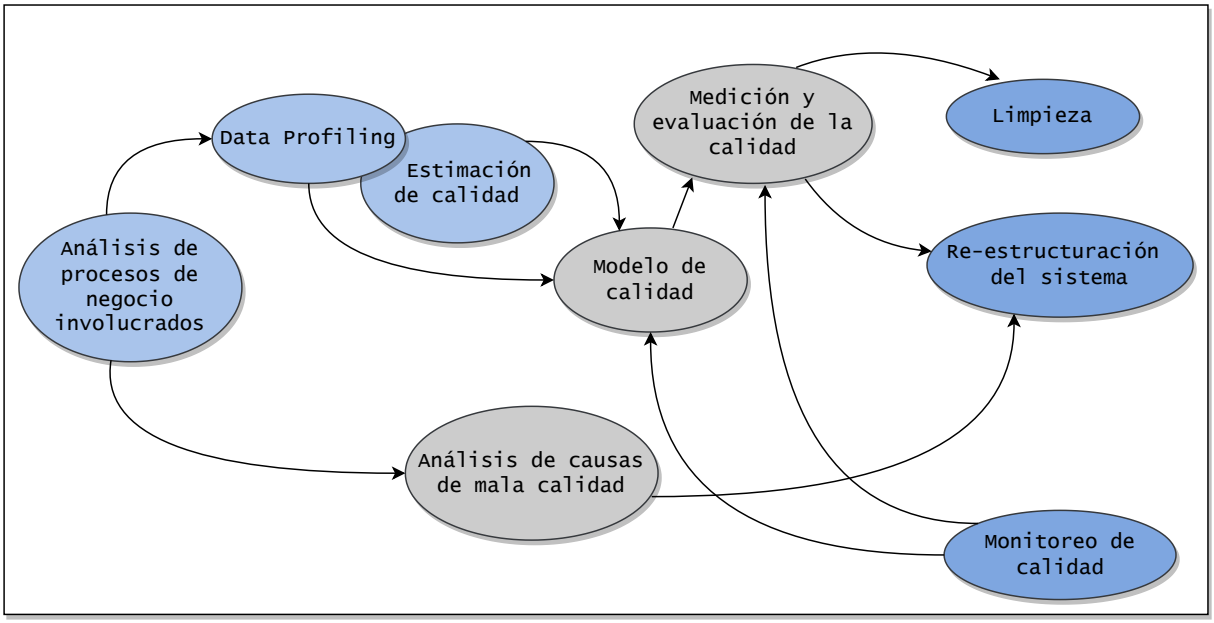


Figura 2.2: Tareas de la gestión de calidad de datos. Extraído de [2].

En primer lugar, se deben analizar los procesos de negocios involucrados; es decir, aquellos procesos que generan, procesan y trabajan con los datos. En segundo lugar, se debe obtener un conocimiento estadístico de las características de los datos, también conocido como *data profiling*. En las siguientes secciones se definirá en detalle esta tarea.

El tercer paso consiste en crear el modelo de calidad de datos que es la guía del trabajo. En el modelo de calidad de datos se definen las dimensiones necesarias que caracterizan el conjunto de datos a tratar. Además, se definen qué datos se priorizarán y cómo se miden.

Luego de definir el modelo de calidad de datos y tener identificados los procesos de negocios involucrados, se realiza un análisis de las causas de la mala calidad de datos. También se puede realizar la medición (ejecutando métricas definidas en el modelo de calidad) y evaluación de los datos para visualizar su estado. Es decir, identificar que tan bien o mal se encuentran.

Otra de las tareas de la gestión de calidad de datos es la limpieza de los datos. La limpieza de los datos, también conocido como *data cleaning* se refiere a la corrección de los problemas de calidad que se tienen en los datos. Esto refiere a la eliminación de datos incorrectos, corruptos, duplicados o incompletos, entre otros [26]. Luego, se realiza la reestructuración del sistema. Esta tarea consiste en corregir los errores detectados en la etapa de análisis. Es decir, corregir las causas de los errores realizando cambios en los procesos.

Por último, y en paralelo a esta tarea se realiza el monitoreo de la calidad de datos. Es necesario contar con esta tarea ya que pueden surgir cambios debido a nuevos requerimientos de calidad de datos o la existencia de nuevos problemas. Para esto, se debe volver a medir y evaluar la calidad existente. En muchos casos, de acuerdo al resultado que arrojen la medición y evaluación se debe modificar el modelo de calidad de manera tal que se adapte al conjunto de datos. La decisión sobre qué tan seguido se debe realizar el control depende del uso que se le dé a los datos y de que tan seguido se modifiquen los procesos de la organización.

La gestión de calidad de datos involucra el conjunto de tareas mencionadas anteriormente. Cabe destacar que se deberá definir un modelo de calidad de datos para cada conjunto de datos, ya que en una organización pueden existir varios conjuntos de datos [25].

2.3.1. Framework y metodología

Uno de los problemas que presenta la gestión de la calidad de datos es que, con las técnicas tradicionales se tiene un cubrimiento limitado de datos y pueden no detectar ciertos errores en los mismos. Una forma de resolver esto es contando con metodologías que permitan sistematizar la gestión de calidad de datos [27][1]. A grandes rasgos, una metodología de calidad de datos se define como un conjunto de pautas y técnicas que, a partir de cierta información de entrada, define un proceso de uso de la información para evaluar y mejorar la calidad de los datos.

Por otro lado, un framework es un conjunto de herramientas que se utiliza para abordar un problema puntual. Los frameworks enfocados en calidad de datos se utilizan en un dominio específico y su función es evaluar y mejorar la calidad de los datos. En el Capítulo 3 se presenta un análisis de distintos frameworks y metodologías enfocados en calidad de datos, y que oficiarán de punto de partida para la construcción del framework presentado en el Capítulo 4.

2.3.2. Algunas herramientas utilizadas en la gestión de calidad de datos

Resulta beneficioso contar con herramientas que cooperen con la gestión de la calidad de datos, permitiendo su evaluación y posterior mejora. En esta sección se introducen herramientas y técnicas que pueden ser utilizadas en distintos momentos de la gestión de la calidad de datos, y que persiguen objetivos diferentes. De acuerdo con la bibliografía relevada, a continuación, se presentan las más utilizadas.

Modelo de madurez

Según Mark C. Paulk *et al.* [28], el concepto de madurez radica en la medida en que un proceso se define, gestiona, mide, controla y es eficaz de forma explícita. Por lo tanto, los modelos de madurez conforman una clase de modelos que se enfocan en los procesos de cambio y desarrollo, y que están relacionados con los sistemas de información y las organizaciones afines.

Un modelo de madurez define un conjunto organizado de atributos o características que sirven para describir distintos aspectos de madurez en un dominio [29]. El principal objetivo de un modelo de madurez es permitir a las organizaciones medir y ubicar diferentes aspectos donde realizar mejoras. Para esto, los modelos de madurez se componen de distintos niveles. Un nivel de madurez consiste en el cumplimiento de prácticas específicas y genéricas para un conjunto de áreas de proceso [30]. Por ende, mejorar el nivel de madurez implica mejorar el desempeño general de la organización.

Data Cleaning

Es una técnica que consiste en implementar estrategias de prevención de errores. “A medida que se identifican patrones de errores, los procedimientos de recopilación e ingreso de datos deben adaptarse para corregir esos patrones y reducir errores futuros” [31]. Esta técnica se conforma por cuatro etapas que son: detección, diagnóstico, tratamiento y documentación.

La detección implica buscar sistemáticamente características sospechosas en cuestionarios de evaluación, bases de datos o conjuntos de datos de análisis. Las etapas de diagnóstico y tratamiento requieren un análisis exhaustivo de todos los tipos y fuentes de errores posibles durante los procesos de recopilación e ingreso de datos. Por último, la documentación de los cambios oficia como auditoría de los errores detectados, alteraciones, adiciones y comprobación de errores. Dicha documentación, también permite volver al valor original en caso de ser necesario [31].

Data Profiling

Data Profiling (perfilado de datos en español) es el proceso de examinar una fuente de datos con el objetivo de obtener información de las características de los datos existentes [32]. Algunos de los resultados más simples son estadísticas tales como la cantidad de nulos o valores distintos en una columna para una determinada tabla. En resultados más complejos se pueden obtener estadísticas relacionadas con dependencias o restricciones funcionales. En general, el perfilado de datos se utiliza como insumo para realizar otras tareas en diversas áreas. A continuación, se presentan algunos ejemplos de uso [32].

- **Optimización de consultas:** la mayoría de los manejadores de base de datos realizan un perfilado de los datos incluidos en las tablas y columnas que se consultan. Con esto, obtienen estadísticas que permiten, entre otras cosas, estimar el costo de un plan de consulta[2].
- **Limpieza de datos:** el perfilado de datos ayuda a descubrir errores de formato, inconsistencias o valores faltantes. En base a esta información se puede implementar un proceso de limpieza.
- **Integración de datos:** en algunos casos se requiere integrar datos de fuentes desconocidas. En ese contexto, puede ser de utilidad explorar las características de los datos tales como cantidad de datos, tipos de datos presentes o la semántica de las columnas y tablas.
- **Analítica de datos:** la mayoría de los análisis de datos cuentan con un perfilado de datos previo. Esto ayuda al analista a entender la naturaleza de los datos.

[2]Un plan de consulta es la secuencia de pasos que realiza el manejador de base de datos para resolver una consulta [33].

Capítulo 3

Análisis de la bibliografía relevante

Este capítulo tiene como objetivo exponer el análisis realizado durante la etapa de investigación de este proyecto. En la Sección 3.1 se presenta el concepto de metodología, mientras que en la Sección 3.2 se introduce el concepto de framework. En la Sección 3.3 se muestran las metodologías y frameworks de uso general que fueron revisados. Luego, en la Sección 3.4 se realiza una comparación de los componentes más relevantes de las metodologías y frameworks de calidad de datos, según las consideraciones de sus autores. Por último, en la Sección 3.5 se presenta un resumen del capítulo.

Aunque tienen significados distintos, los conceptos de metodología y framework tienden a confundirse. En el contexto de este proyecto es necesario entender la diferencia que existe entre ambos. A continuación, se presentan ambas definiciones.

3.1. Metodología

Según la Real Academia Española (RAE) [34], una metodología es un conjunto de métodos que se siguen en una investigación científica o en una exposición doctrinal. Batini *et al.* [16], define una metodología de calidad de datos como un conjunto de pautas y técnicas que, a partir de cierta información de entrada, define un proceso de uso de la información para evaluar y mejorar la calidad de datos.

3.1.1. Taxonomía de metodologías de calidad de datos

Existen varios criterios para clasificar las metodologías de calidad de datos. Batini *et al.* [16] propone la siguiente taxonomía:

- **Impulsado por información vs. impulsado por procesos:** Las estrategias impulsadas por la información se centran en el uso de fuentes de información que son exclusivas para mejorar la calidad de la misma. En cambio, en las estrategias impulsadas por procesos, el proceso de producción de información se analiza y posiblemente se modifica para poder identificar y eliminar las causas fundamentales de los problemas de calidad.

- **Evaluación vs. mejora:** Las actividades de las metodologías de evaluación y mejora están estrechamente relacionadas. Al disponer las mediciones de calidad de datos es posible seleccionar las técnicas que luego serán aplicadas para su mejora. Como consecuencia, el límite entre las metodologías de medición y mejora a veces es impreciso. En este informe, se utilizará el término medición cuando se trate de medir los valores de un conjunto de dimensiones de calidad de datos en una fuente de información. Por otro lado, se empleará el término evaluación cuando dichas mediciones se utilicen para analizar la calidad de los datos de la organización.

- **Propósito general vs. propósito especial:** Una metodología de propósito general cubre un amplio espectro de fases, dimensiones y actividades. Por otro lado, una metodología de propósito especial se enfoca en una actividad específica, en un dominio de información específico, o en dominios de aplicación específicos.

- **Intraorganizacional vs. Interorganizacional:** En las metodologías de carácter intraorganizacional, la actividad de medición y mejora involucra a una organización específica o un sector específico de la organización. En cambio, para las metodologías de tipo interorganizacional, las actividades están diseñadas para un grupo de organizaciones que cooperan con un objetivo común.

3.2. Framework

En la literatura existen varias definiciones de framework. Según Othman [35], un framework es una guía que se utiliza para establecer dominios, objetivos, procesos con entradas y salidas, roles y responsabilidades. Además, tiene como objetivo establecer una forma de administrar una organización. Por otro lado, Van Heesch [36] menciona la norma ISO 2011 para definir un framework como un conjunto de prácticas que son desarrolladas y ejecutadas para resolver una problemática específica. Es una forma de estandarizar procesos y orientarlos al cumplimiento de objetivos.

De lo expresado, se desprende que un framework orientado a la gestión de la calidad de datos tiene como objetivo contribuir a la administración de las organizaciones con el fin de mejorar la calidad de los datos que se generan y/o utilizan. Por lo tanto, un framework brinda pautas que sirven como guía para resolver un problema determinado, mientras que una metodología propone la ejecución de un proceso basado en un conjunto de métodos.

3.3. Revisión de metodologías y frameworks

En general, tanto los frameworks como las metodologías se definen para un dominio en particular. En la actualidad, no existe un estándar para la definición de un framework de uso genérico enfocado en la gestión de la calidad de datos. Por este motivo, es necesario realizar un relevamiento de los trabajos existentes, con el fin de establecer un punto de partida desde el cual definir un framework de propósito general para la gestión de la calidad de datos.

Según Cichy, que realizó una revisión proporcionando frameworks de gestión de calidad de datos [37], un framework con estas características debe proporcionar:

- Definición de la calidad de los datos y sus características más relevantes.
- Actividades para la evaluación de la calidad de los datos.
- Actividades para la mejora de la calidad de los datos.

En la **Tabla 3.1** se presentan cuatro metodologías y dos frameworks. Los mismos están enfocados en la gestión de calidad de datos y cumplen con los criterios mencionados anteriormente. De aquí en adelante, serán mencionados utilizando el acrónimo indicado en dicha tabla.

Acrónimo	**Descripción**	**Año**
HDQM	A Data Quality Methodology for Heterogeneous Data	2011
CDQ	Comprehensive Methodology for Data Quality Management	2006
AIMQ	A methodology for information quality assessment	2002
TDQM	Total Data Quality Managment	1998
FW-AGESIC	Framework para la Gestión de Calidad de Datos en Gobierno Digital	2020
DQMP	Data Quality Framework for the Estonian Public Sector	2017

Tabla 3.1: Frameworks y metodologías analizados en el estado del arte.

Cada metodología y framework propone un proceso para gestionar la calidad de los datos donde se definen actividades para las etapas de evaluación y mejora de la calidad de datos. En algunas propuestas, además del proceso de gestión, se incluyen otros componentes como, por ejemplo, un modelo de madurez o un equipo responsable de la aplicación del framework. En otras propuestas, las etapas de evaluación y mejora de calidad de datos se consideran componentes independientes. En esta sección se presentarán los componentes considerando la visión de los autores.

En la **Tabla 3.2** se exhiben los principales componentes de cada metodología y framework.

Acrónimo	Componentes
HDQM	Reconstrucción del estado, evaluación/medición de calidad de los datos, selección del proceso de mejora [38].
CDQ	Reconstrucción del estado, evaluación y elección del proceso de mejora [39].
AIMQ	Modelo de calidad de datos, evaluación y elección de técnicas de valoración e interpretación de la información [40].
TDQM	Definir, medir, analizar y mejorar (Ciclo TDQM) [14].
FW-AGESIC	Marco teórico, proceso de gestión de calidad de datos, marco conceptual, caso de estudio, modelo de calidad de datos de referencia y recursos de soporte [3].
DQMP	Modelo de calidad de datos, modelo de madurez de la calidad de datos y proceso de gestión de la calidad de datos [4].

Tabla 3.2: Principales componentes de las metodologías y frameworks analizados.

3.3.1. Metodologías

A continuación, se presentan en detalle las cuatro metodologías mencionadas en la **Tabla 3.1**.

HDQM: La metodología HDQM [38] fue propuesta en el año 2011 y es una de las pocas metodologías considerada heterogénea ya que puede aplicarse sobre distintos tipos de datos. Para la evaluación y mejora de dichos datos son tomados en cuenta, datos estructurados, semiestructurados y no estructurados. Para esto, HDQM propone mapear los diferentes tipos de datos a una representación conceptual común. Luego, evalúa la calidad de éstos considerando una única representación. HDQM está formado por tres componentes: reconstrucción del estado, evaluación y medición de los datos, y por último, selección del proceso de mejora de los datos. A continuación, se presentan las actividades de cada componente.

- **Reconstrucción del estado:** este componente tiene como objetivo reconstruir todo el conocimiento relevante sobre la organización, es decir, sus procesos, recursos, entidades, entre otros.
- **Evaluación/medición de los datos:** este componente tiene como propósito obtener una

evaluación cuantitativa de los problemas de calidad de datos presentes en la organización. Consiste en dos actividades: clasificación de los datos obtenidos en la etapa anterior y la medición de su calidad.

- **Selección del proceso de mejora:** este componente comprende tres actividades: definición de requisitos de calidad de datos, selección de la actividad de mejora de la calidad de datos y, por último, la elección y evaluación del proceso de mejora óptimo.

Los tres componentes de la metodología, sus entradas y salidas se muestran en la **Figura 3.1**. Esta figura representa el despliegue de la metodología HDQM de una manera simplificada y secuencial. La misma no describe la posible retroalimentación entre los componentes.

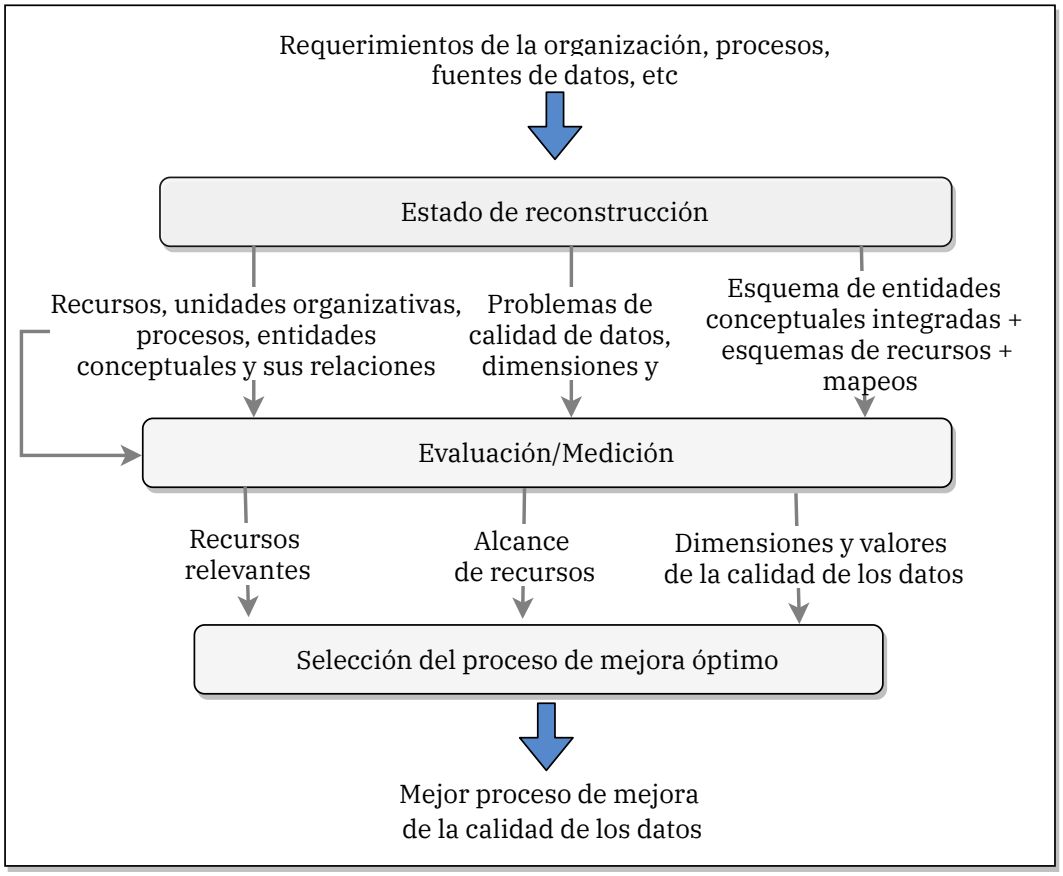


Figura 3.1: Esquema de componentes, entradas y salidas de la metodología HDQM. Extraído y adaptado de [38].

CDQ: La metodología CDQ [39] fue propuesta en el año 2006. Proporciona pautas y técnicas para analizar los aspectos relevantes relacionados con la calidad de datos en el ámbito empresarial. La metodología se centra en integrar de forma genérica, etapas, técnicas y herramientas propuestas en otros frameworks conocidos como TDQM [14], TIQM [41] e Istat [42]. Es una metodología flexible ya que apoya al usuario en la selección de las técnicas y herramientas más adecuadas en cada una de sus etapas.

Al igual que la metodología HDQM (conocida por ser una extensión de esta metodología), los tres principales componentes de CDQ son: reconstrucción del estado, evaluación, y elección del proceso de mejora óptimo. Cada uno de ellos tiene un objetivo específico y un conjunto de técnicas a aplicar.

- **Reconstrucción del estado:** este componente consiste en modelar y documentar las relaciones entre unidades organizativas, procesos, servicios y fuentes de datos.
- **Evaluación:** este componente tiene como objetivo medir la calidad de datos actual. Para ello, es necesario seleccionar un subconjunto de dimensiones de calidad relevante y métricas relacionadas. Luego, se deben implementar y aplicar las métricas de calidad a los datos para proporcionar una evaluación cuantitativa de los problemas de calidad identificados en el paso anterior.
- **Mejora:** este componente consiste en definir un proceso de mejora que tiene como objetivo resolver los problemas de calidad de datos identificados.

A modo de facilitar la comprensión de la metodología CDQ, la **Figura 3.2** presenta un esquema de los componentes, entradas y salidas de la misma.

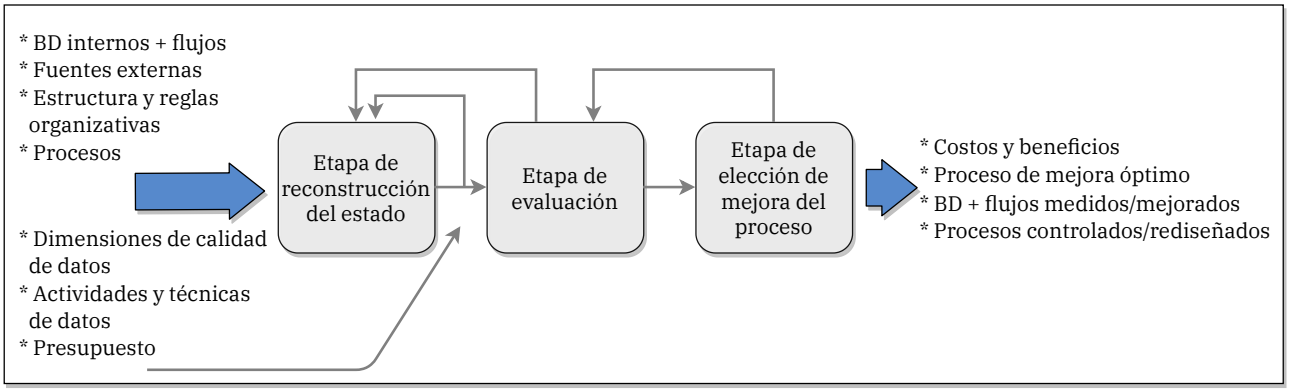


Figura 3.2: Esquema de componentes, entradas y salidas de la metodología CDQ. Extraído y adaptado de [39].

AIMQ: La metodología AIMQ [40] se propuso en el año 2002 y fue ideada para la evaluación de la calidad de la información centrada en la evaluación comparativa. Es una metodología integral que se enfoca en satisfacer los requisitos de quienes administran la información, así como de quienes hacen uso de ésta. Dependiendo de cómo sea valorada la información por administradores y usuarios, se utilizan cuatro cuadrantes para clasificar los datos. A continuación, se describen los tres componentes de ésta metodología:

- **Definir el modelo de calidad de datos:** el primer componente tiene como fundamento la receptividad de la información, es decir, la manera en que es percibida. Es una matriz que tiene cuatro cuadrantes, dependiendo de si la información se considera un producto o un servicio y de si las mejoras pueden evaluarse frente a una especificación formal o una expectativa del cliente.
- **Evaluación de la calidad de la información:** este componente realiza un cuestionario para medir la calidad de la información. Para ello, se establecen las variables que integran la calidad de los datos y se extraen las dimensiones, para luego medir la calidad de los mismos sujetos a la opinión de los usuarios. Con esto como base, se analiza si las mejoras efectuadas al sistema pueden valorarse a través de descripciones formales o desde la perspectiva del cliente.
- **Elección de técnicas:** el tercer componente consiste en dos técnicas de valoración e interpretación de la información. La primera de ellas captura la información que se obtuvo al aplicar el cuestionario de medición y la segunda técnica mide el trayecto entre una evaluación y otra.

La **Figura 3.3** representa un esquema de los componentes de la metodología presentada.

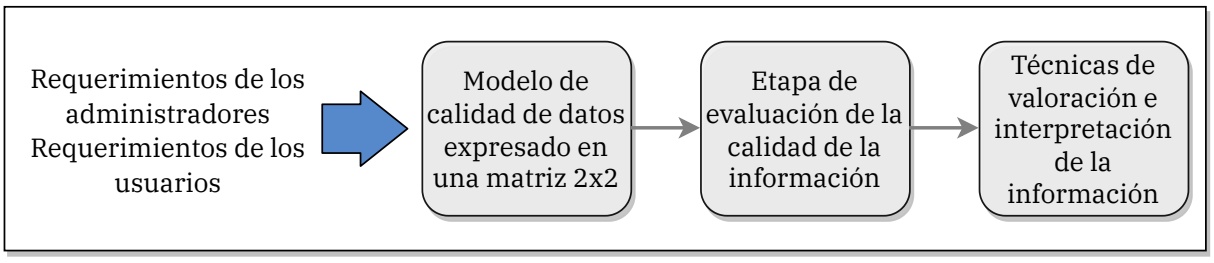


Figura 3.3: Esquema de los componentes de la metodología AIMQ.

TDQM: Es una metodología propuesta por Wang [14] en el año 1998 y tiene como objetivo la entrega de productos de información de buena calidad para los consumidores de información. Según Wang, un producto de información (PI) es el resultado o salida de un sistema de información que genera valor para el consumidor [43].

Para poder aplicar esta metodología, es fundamental que las organizaciones traten la información como un producto que se mueve a través de un sistema de fabricación, muy similar a un producto físico, pero con las ventajas y limitaciones propias de la naturaleza de un PI. Esta metodología propone un ciclo de mejora continua, compuesto por cuatro etapas:

- **Definición:** tiene como objetivo identificar los requisitos de calidad de datos de la organización, así como también, las dimensiones de calidad de datos asociadas. Por otro lado, también es necesario definir el sistema de producción de información, es decir, el conjunto de personas, equipos y programas involucrados en la transformación de la información de la organización.
- **Medición:** se encarga de producir las métricas de calidad de datos que brindan retroalimentación a la gestión de la calidad de los datos.
- **Análisis:** identifica las principales causas de los problemas de calidad de datos y calcula el impacto de la información de mala calidad.
- **Mejora:** diseñar actividades de mejora de la calidad, identificando las áreas clave para las mejoras.

La **Figura 3.4** representa a grandes rasgos, el esquema de la metodología TDQM.

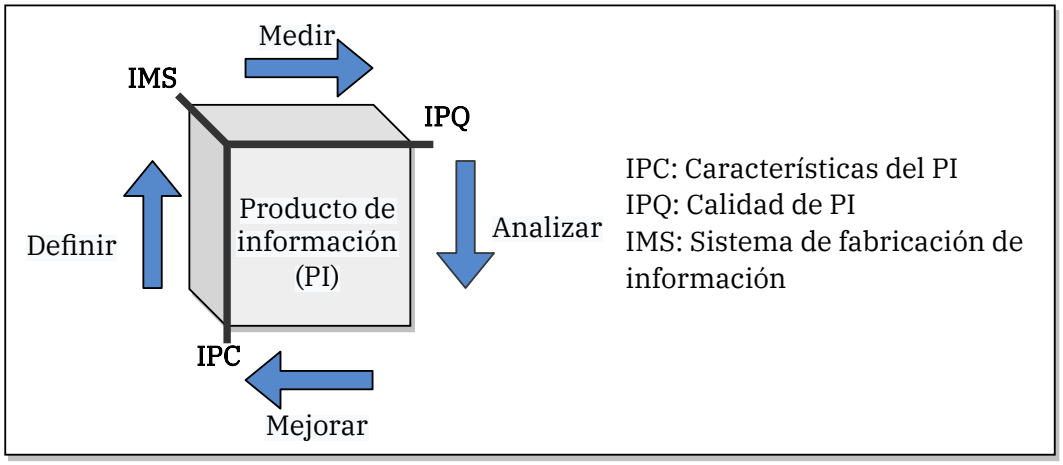


Figura 3.4: Esquema de etapas de la metodología TDQM. Extraído y adaptado de [14].

Por último, al aplicar TDQM, una organización debe tener en cuenta lo siguiente [43]:

- **Enseñar a realizar una evaluación de la calidad de datos aplicando la metodología:** para poner en práctica la metodología, se debe dar la orientación y capacitación necesaria. Se debe comprobar con el equipo de calidad de datos la efectividad de implementar la metodología y dar a conocer sus problemas prácticos y los posibles puntos débiles.
- **Establecer un equipo de calidad de datos que utilizará TDQM:** se recomienda formar un equipo de personas idóneas en la temática para aumentar las probabilidades de éxito al implementar la metodología.
- **Definir un producto de información en términos de negocio:** cuando un atributo específico del producto de información está mejorado con TDQM, en su primera fase de definición se describen los datos del producto con mayor claridad.
- **Institucionalizar la mejora continua de los productos de información:** una vez que la aplicación de la metodología haya finalizado, la organización debe institucionalizar el método diseñado y el equipo de calidad de datos para asegurar una mejora continua de sus datos.

3.3.2. Frameworks

A continuación, se presentan en detalle los dos frameworks mencionados en la **Tabla 3.1**.

FW-AGESIC: El framework propuesto por AGESIC [1] se creó en Mayo de 2020 con el objetivo de contribuir a la sistematización de la gestión de la calidad de datos en organizaciones vinculadas al gobierno digital en Uruguay. Su objetivo es mejorar la calidad de los datos que se generan y/o utilizan en el territorio uruguayo.

Los principales componentes del framework son: un marco teórico, un marco conceptual y un modelo de calidad de datos de referencia que provee un conjunto extensible e instanciable de elementos de calidad de datos. Este modelo de referencia es una guía para la definición de modelos de calidad para escenarios de trabajo específicos. Además, este framework cuenta con un proceso para la gestión de calidad de datos y recursos de soporte [3]. El proceso de gestión de calidad de datos está compuesto por siete etapas. Las mismas se detallan a continuación:

[1] https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/

- **Caracterizar el escenario:** el objetivo es identificar los elementos relevantes del escenario de trabajo. Se debe considerar los puntos de vista técnico, de negocio y de calidad de datos.
- **Examinar los datos objetivos:** el objetivo de esta etapa es conocer las características de los datos y realizar una primera aproximación de su calidad. También se busca detectar problemas de calidad de datos que puedan derivar en nuevos requisitos de calidad.
- **Definir estrategia de gestión de calidad:** el objetivo es definir la estrategia para gestionar la calidad de datos en el escenario de trabajo, en base a pasos que abordan un conjunto de requisitos de calidad de datos.
- **Definir modelo de calidad de datos:** este modelo se construye utilizando el modelo de calidad de datos de referencia. Debe construirse en base a los requisitos de calidad de datos identificados en la estrategia definida en la etapa anterior.
- **Medir y evaluar calidad de datos:** consiste en realizar mediciones utilizando las métricas y métodos definidos en el modelo de calidad. Se debe evaluar la calidad de los datos en base a los perfiles y reglas de evaluación definido.
- **Determinar causas de los problemas de calidad:** esta etapa tiene como objetivo determinar las causas de los problemas de calidad identificados en las mediciones y evaluaciones realizadas en la etapa anterior. Para esto se pueden tomar como insumo tanto los resultados de las mediciones y evaluaciones, como los resultados obtenidos de la caracterización del escenario.
- **Definir, ejecutar y evaluar el plan de mejora:** esta etapa tiene como objetivo la definición de un plan de mejora para abordar los problemas de calidad detectados, así como también, su ejecución y posterior evaluación.

La **Figura 3.5** representa el proceso de gestión de calidad de datos del framework.

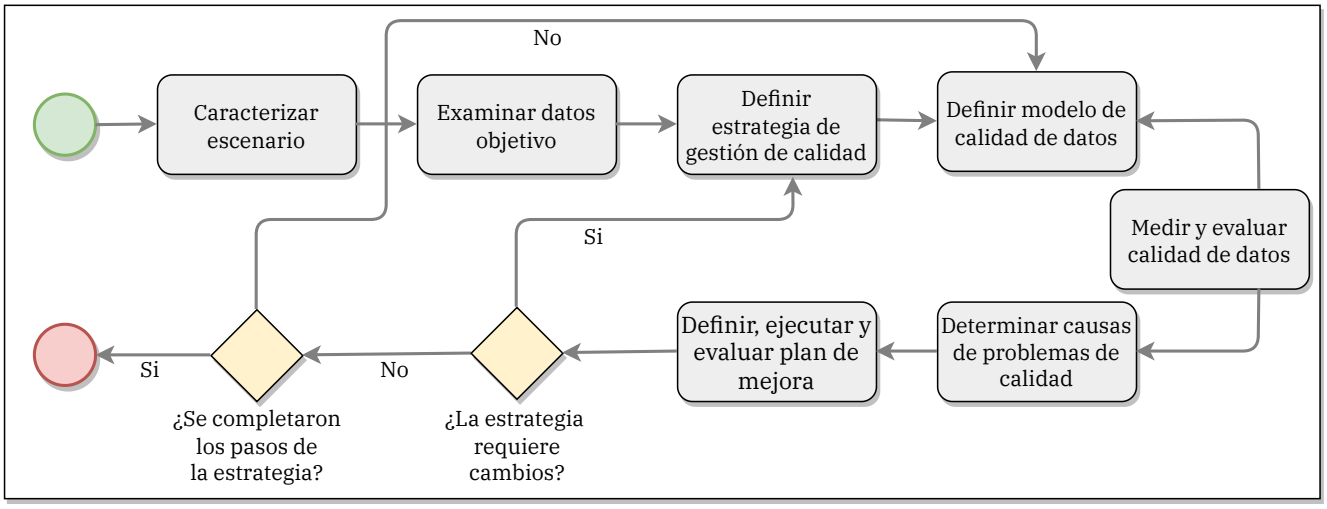


Figura 3.5: Proceso de gestión de calidad del framework propuesto por AGESIC. Extraído y adaptado de [3].

DQMP: El framework DQMP [4] fue creado para el gobierno de Estonia en el año 2017, con el objetivo de mejorar la calidad de datos en el sector público. Los componentes principales del framework son un modelo de calidad de datos, un modelo de madurez, un conjunto de actividades a nivel estatal y un proceso de gestión de la calidad de los datos.

- **Modelo de calidad de datos:** conjunto de características que se utilizan para especificar los requisitos y poder evaluar la calidad de los datos.
- **Modelo de madurez:** es una herramienta para que el propietario de un sistema de información evalúe la madurez de la organización con respecto a la calidad de los datos y planifique las mejoras.
- **Proceso de gestión de calidad de datos:** conjunto de actividades para avanzar hacia una mejor calidad de los datos dentro de una organización. Esto último se conoce como proceso de gestión de calidad de los datos que se basa en el ciclo OPDC. Este ciclo se compone por las siguientes etapas: observar, planificar, hacer, verificar y ajustar/actuar.

Para que este framework funcione en todo el sector público, se proporciona un conjunto de actividades a nivel estatal con el objetivo de mejorar la calidad de los datos en dicho sector. La **Figura 3.6** muestra los componentes propuestos por el framework DQMP.

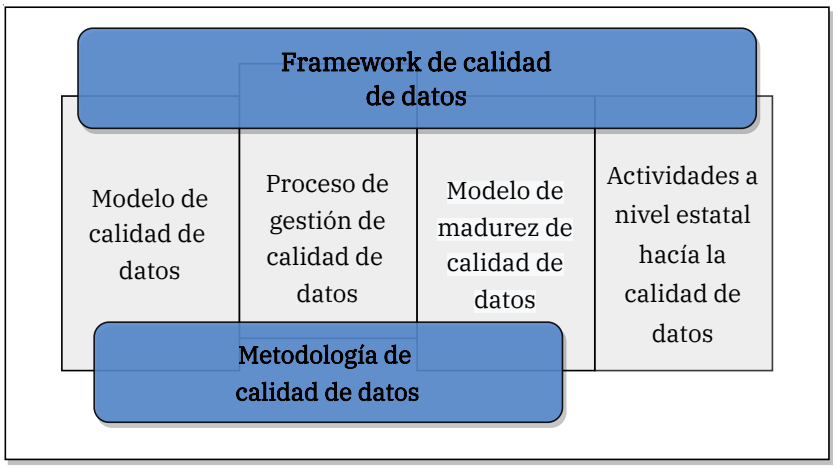


Figura 3.6: Componentes del framework DQMP propuesto por el gobierno de Estonia. Extraído y adaptado de [4].

3.4. Componentes relevantes de las metodologías y frameworks

Las metodologías y frameworks analizados cuentan con componentes, actividades y/o herramientas para gestionar la calidad de datos. A continuación, se presentan los componentes de las metodologías y framework analizados que son relevantes para los objetivos de este proyecto.

3.4.1. Proceso de gestión para la mejora de la calidad de datos

El proceso de gestión de calidad de datos es un componente común a las metodologías y frameworks revisados. Sin embargo, existen diferencias en la nomenclatura por parte de varios autores. Algunas metodologías y frameworks incluyen la evaluación y mejora como componentes independientes, mientras que en otros, forman parte del proceso de gestión de calidad de datos. En este proyecto, se considerará a la evaluación y mejora como etapas del proceso de gestión.

El proceso de gestión para la mejora de la calidad de datos propone la ejecución sucesiva de etapas con el fin de gestionar la calidad de los datos. Algunos de los autores consultados difieren en la cantidad de etapas y las actividades a realizar dentro de cada una de ellas. Sin embargo, todos coinciden en incluir dos etapas consideradas fundamentales: la etapa de evaluación y la etapa de mejora de calidad de datos.

Etapa de evaluación de la calidad de datos

En esta sección se presenta el análisis de la etapa de evaluación de la calidad de datos propuesto en las metodologías y frameworks revisados en este proyecto. Esta etapa está formada por una serie de actividades que involucran distintos roles dentro de una organización.

Según Batini *et al.* [16], la etapa de evaluación de calidad de datos tiene como objetivo brindar un diagnóstico preciso del estado del sistema de información con respecto a las cuestiones de calidad de datos. Esto es, presentar las mediciones de calidad de datos de las fuentes de información, presentar los costos para la organización debido a la calidad actual y, por último, brindar una comparación con los niveles de calidad de datos considerados aceptables. Dichos niveles están basados en la experiencia o en un documento con sugerencias de mejora. Por otro lado, Batini plantea que las actividades más comunes en la etapa de evaluación son:

- **Análisis:** examina las fuentes de datos e información, esquemas y metadatos disponibles en las organizaciones. Luego, realiza entrevistas para alcanzar una comprensión completa de la información, de las reglas de arquitectura y gestión relacionadas.
- **Análisis de los requisitos de calidad de datos:** obtiene la opinión de los usuarios y administradores de la información con el fin de identificar problemas de calidad y establecer nuevos objetivos de calidad.
- **Identificación de áreas críticas:** se seleccionan las fuentes de información y los flujos más relevantes para ser evaluados cuantitativamente.
- **Modelado de procesos:** proporciona un modelo de los procesos que producen o procesan la información.
- **Medida de calidad:** selecciona las dimensiones de calidad que son afectadas por los problemas de calidad identificados en la etapa de análisis de requisitos de calidad de datos para definir las métricas correspondientes.

En la metodología HDQM, la etapa de evaluación consta de dos grandes actividades, la clasificación de los datos relevados y la medición de la calidad de los datos. El objetivo de la primera es establecer la viabilidad y riesgos para la actividad de mejora. Por otra parte, la actividad de medición de la calidad de datos tiene como objetivo obtener una evaluación cuantitativa de los problemas de calidad identificados previamente. Para esto, es necesario seleccionar las dimensiones de calidad de datos relevantes que aplican y las métricas relacionadas. Las dimensiones se miden aplicando las métricas asociadas, y con esto se proporciona una evaluación cuantitativa

de la calidad [38]. En la **Figura 3.7** se presenta un diagrama con las actividades mencionadas anteriormente para esta etapa.

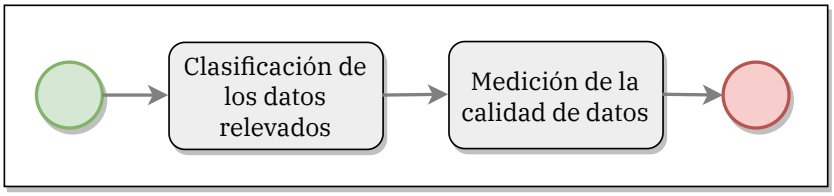


Figura 3.7: Actividades para la etapa de evaluación de la metodología HDQM [38].

En la metodología CDQ, previo a la etapa de evaluación, se realiza una reconstrucción del estado con el objetivo de obtener una imagen completa del uso de los datos y el flujo entre los mismos. A diferencia de otras metodologías, CDQ enfatiza la importancia de obtener los requisitos de la organización previo a la etapa de evaluación.

La etapa de evaluación de CDQ consta de dos pasos. Por un lado, la identificación de problemas de calidad de datos que involucra tanto a los usuarios internos como externos de la organización. El resultado del primer paso de la etapa de evaluación es la base para el análisis de los procesos identificados en la fase de reconstrucción y la identificación de las causas de la mala calidad de los datos. El segundo paso de la evaluación consiste en la medición de la calidad de datos y tiene como objetivo obtener una evaluación cuantitativa de los problemas de calidad que existen. Para esto se selecciona un subconjunto de dimensiones relevantes de calidad de datos y las métricas asociadas. Luego, se aplican dichas métricas a los datos para proporcionar una evaluación cuantitativa de los problemas de calidad identificados en el paso anterior [39]. En la **Figura 3.8** se presenta un diagrama con las actividades mencionadas anteriormente para esta etapa.

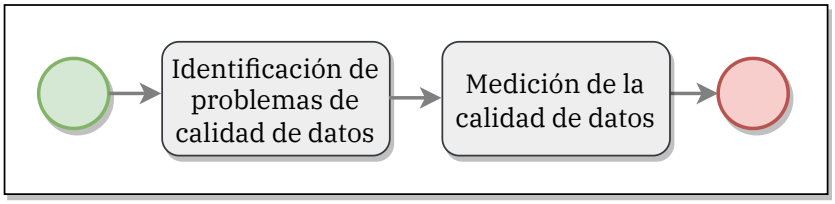


Figura 3.8: Actividades para la etapa de evaluación de la metodología CDQ.

En la metodología AIMQ, la calidad de datos se evalúa mediante cuestionarios que fomentan la investigación sobre el éxito o no, de los sistemas de información. Luego, se analizan los resultados obtenidos y se priorizan tareas, de forma tal de asignar los recursos necesarios para realizar la mejora de la calidad de datos [40].

Algunas metodologías, como es el caso de TDQM, no proporcionan etapas formales enfocadas en la evaluación de calidad de datos [14]. En este caso, existen actividades que están relacionadas con la evaluación. Por ejemplo, desarrollar las métricas de calidad de datos que fueron identificadas. Luego, estas métricas serán implementadas en un sistema de fabricación de información.

Para la evaluación de la calidad de datos, el framework FW-AGESIC [3] identifica los elementos relevantes del escenario de trabajo. Para esto, se utilizan dos enfoques: uno orientado a los aspectos técnicos y de negocio, y otro orientado a los aspectos de calidad de datos. Esto último se realiza para identificar los requisitos de calidad y los problemas de calidad de datos. Una vez identificado el escenario, se examinan los datos objetivos para conocer sus características y obtener una primera aproximación de su calidad. Luego, se define una estrategia de gestión de calidad de datos que proporciona una serie de pasos a seguir para abordar un conjunto de requisitos y definir el modelo de calidad de datos.

Por último, para evaluar la calidad de los datos se realizan mediciones utilizando las métricas y métodos definidos en el modelo de calidad. También, se evalúa la calidad de los datos en base a los perfiles y reglas de evaluación definidos en el modelo. En la **Figura 3.9** se presenta un breve diagrama que incluye las actividades mencionadas anteriormente para esta etapa.

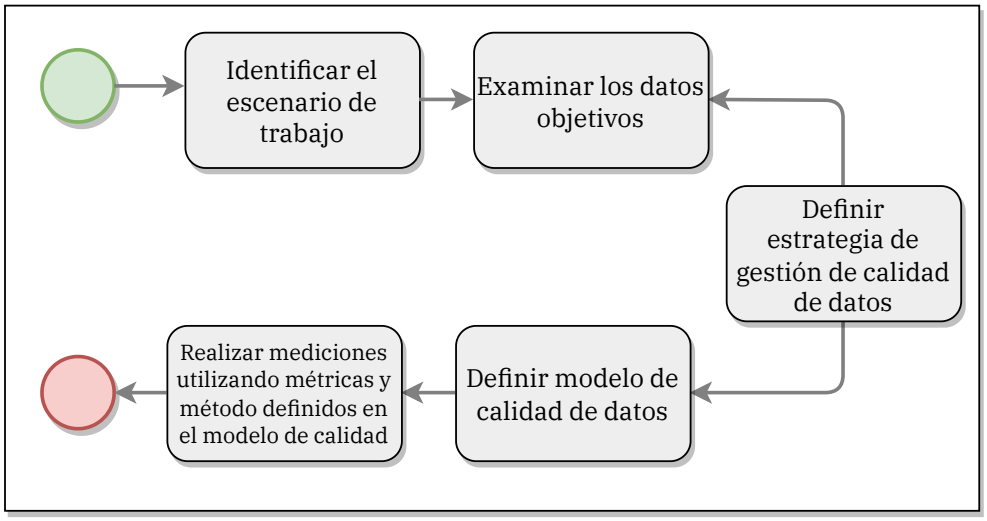


Figura 3.9: Actividades de la etapa de evaluación del framework propuesto por AGESIC [3].

En la etapa de evaluación del framework DQMP [4] se debe observar la situación actual del sistema, para luego evaluar la necesidad de mejora. Para esto, se debe definir el alcance de los datos que se evaluarán. Luego, se debe identificar el área de la organización donde se evaluarán estos datos e identificar a los responsables de cada área. Finalmente, se realizará la evaluación

del nivel actual de madurez de la organización con respecto a la calidad de datos. Esta evaluación proporciona estimaciones de los niveles de optimización de los procesos de la organización para cada factor de madurez y una calificación general del nivel de madurez de la organización. En la **Figura 3.10** se presenta un breve diagrama de las actividades de la etapa mencionada anteriormente.

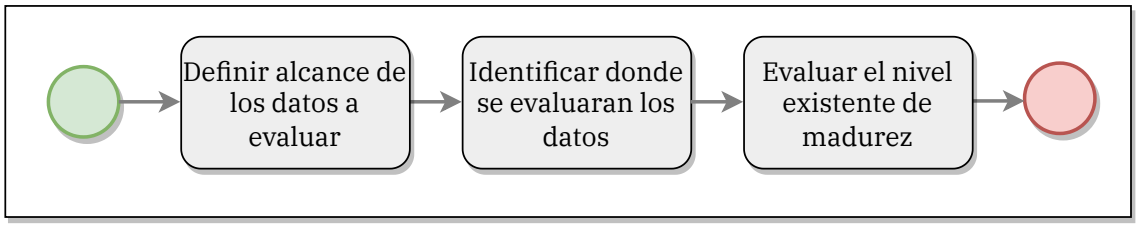


Figura 3.10: Actividades de la etapa de evaluación del framework DQMP.

Etapa de mejora de la calidad de datos

En la mayoría de las metodologías y frameworks de gestión de la calidad de datos, la etapa siguiente a la evaluación consiste en definir acciones para mejorar la calidad de datos. Según Batini *et al.* [16], las actividades de la etapa de mejora se centran en:

- **Evaluación de costos:** estima los costos directos e indirectos de la calidad de los datos.
- **Asignación de responsabilidades:** identifica a los propietarios de la información y define sus responsabilidades en la gestión de los datos.
- **Identificación de las causas de los errores:** identifica las causas de los problemas de calidad.
- **Diseño de soluciones de mejora de la información:** selecciona la estrategia más eficaz y eficiente, y el conjunto de técnicas y herramientas relacionadas para mejorar la calidad de la información.
- **Control de procesos:** define puntos de control en los procesos de producción de información para monitorizar la calidad durante la ejecución del proceso.
- **Rediseño de procesos:** define las acciones de mejora del proceso que pueden generar las correspondientes mejoras de calidad de datos.
- **Gestión de la mejora:** define nuevas reglas organizativas para la mejora de la calidad de la información.

- **Seguimiento de la mejora:** establece actividades de seguimiento de forma periódica que retroalimenten los resultados del proceso de mejora y posibiliten su sintonía dinámica.

A continuación, se realiza una exposición de las principales actividades de mejora de calidad de datos adoptadas por las metodologías y frameworks revisados.

La metodología HDQM propone tres actividades en la etapa de mejora de la calidad de datos. Estas actividades son: la definición de requisitos de calidad de datos, la selección de la actividad de mejora de la calidad de datos y la elección del proceso de mejora óptimo. La primera actividad consiste en establecer los valores objetivos de calidad de datos. La segunda actividad consiste en seleccionar las actividades impulsadas por datos y por procesos que son candidatas a ser elegidas para un proceso de mejora óptimo. Finalmente, en la tercera actividad se elige un proceso que satisfaga la inclusión de todas las actividades necesarias para mejorar el conjunto de dimensiones de calidad de datos. En la **Figura 3.11** se presenta un diagrama con las actividades mencionadas anteriormente.

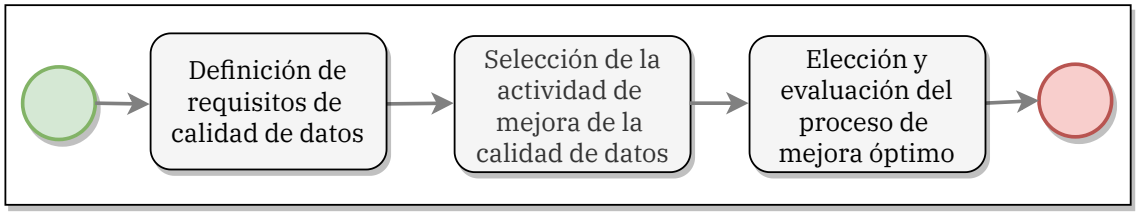


Figura 3.11: Actividades para la etapa de mejora de la metodología HDQM.

En la metodología CDQ se sugiere la elección del mejor proceso para la mejora de calidad de datos. Para realizar dicha elección se proponen cinco actividades. En la primera, se deben definir los requisitos de calidad de datos. En la segunda, se deben seleccionar las actividades de mejora de la calidad de datos. En la tercera, se elige la mejor técnica para las actividades seleccionadas. En cuarto lugar, se elige la técnica de mejora más adecuada para los procesos. Por último, se debe evaluar la efectividad en dichos procesos [39]. En la **Figura 3.12** se presenta un diagrama con las actividades mencionadas anteriormente.

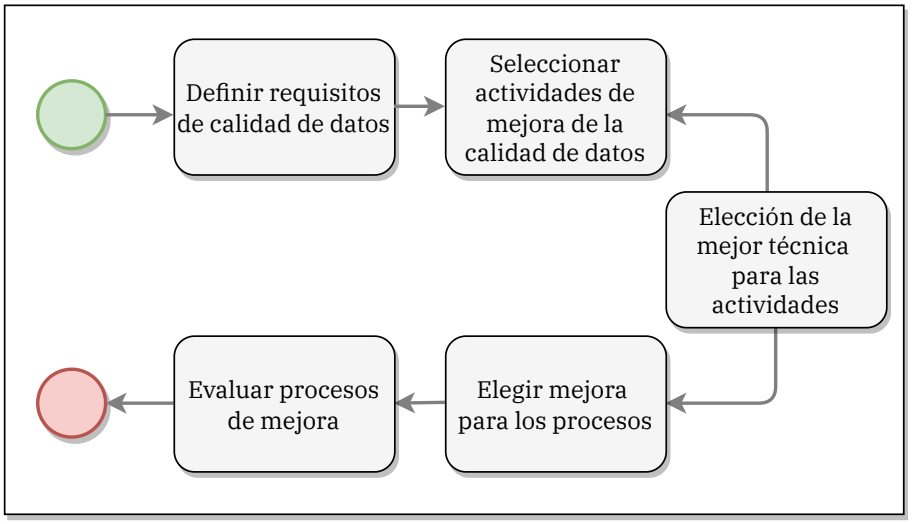


Figura 3.12: Actividades para la etapa de mejora de la metodología CDQ.

La metodología TDQM [14] necesita identificar las áreas claves para la mejora y alineamiento de los flujos de información. Luego, se proporcionan técnicas para mejorar la calidad de los datos.

El framework FW-AGESIC [3], propone en primer lugar, determinar las causas de los problemas de calidad relevados de las mediciones y evaluaciones realizadas en la etapa anterior. Para esto, se pueden tomar como insumo tanto los resultados de las mediciones, como los resultados obtenidos de la caracterización del escenario. Luego, define un plan de mejora para abordar los problemas de calidad detectados, así como la ejecución y posterior evaluación de este plan. En la **Figura 3.13** se presenta un diagrama con las actividades mencionadas anteriormente para esta etapa.

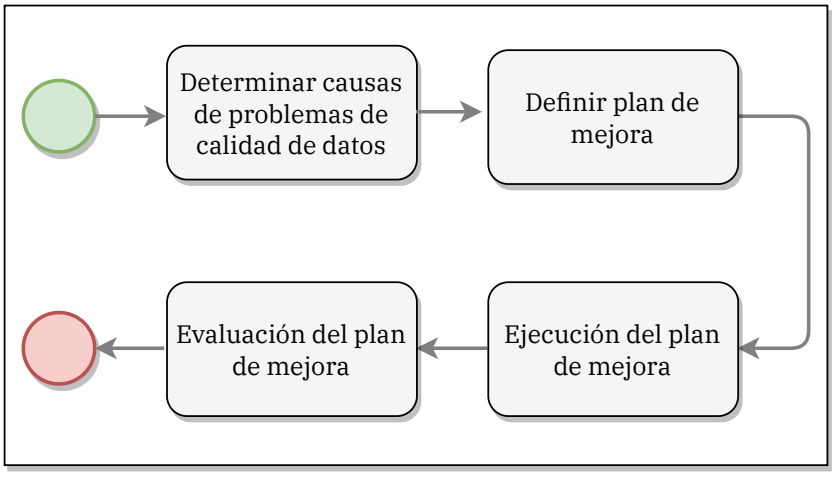


Figura 3.13: Actividades de la etapa de mejora para el framework propuesto por AGESIC [3].

Por último, en el framework DQMP [4], el plan de mejora de la calidad de los datos comprende las actividades necesarias para alcanzar el nivel objetivo de madurez del sistema para la iteración actual. Para cada factor de madurez, el plan identifica cuales fueron evaluadas como falsas, y define actividades para hacer que esto sea verdadero.

3.4.2. Modelo de calidad de datos

Según el estándar ISO 25012 [44] un modelo de calidad de datos representa los cimientos sobre los cuales se construye un sistema para la evaluación de un producto de datos. En un modelo de calidad de datos se establecen las características de calidad que se deben tener en cuenta a la hora de evaluar las propiedades de un producto de datos determinado. La identificación de las dimensiones más relevantes puede considerarse en la etapa de evaluación y sirve como punto de partida para diversas actividades de la etapa de mejora. Esto se observa en muchas de las metodologías y frameworks revisados. Por otro lado, el modelo de calidad depende principalmente del dominio, del framework o metodología aplicada y del tipo de datos de una organización. En la **Tabla 3.3** se presentan las principales dimensiones que se incluyen en cada metodología y framework analizado.

Acrónimo	**Dimensiones**
HDQM	Exactitud, Actualidad [38].
CDQ	Integridad, Actualidad, Relevancia, Flexibilidad [39].
AIMQ	Accesibilidad, Credibilidad, Integridad, Puntualidad, Relevancia [40].
TDQM	Precisión, Objetividad, Credibilidad, Accesibilidad, Relevancia, Oportunidad, Integridad, Coherencia [14].
FW-AGESIC	Exactitud, Consistencia, Completitud, Unicidad, Frescura [3].
DQMP	Precisión, Integridad, Consistencia, Credibilidad, Actualidad, Confidencialidad, No redundancia [4].

Tabla 3.3: Dimensiones de calidad de datos de las metodologías y frameworks analizados.

Por último, en la **Figura 3.14** se puede observar que las dimensiones que más se mencionan son: integridad, actualidad, relevancia y credibilidad.

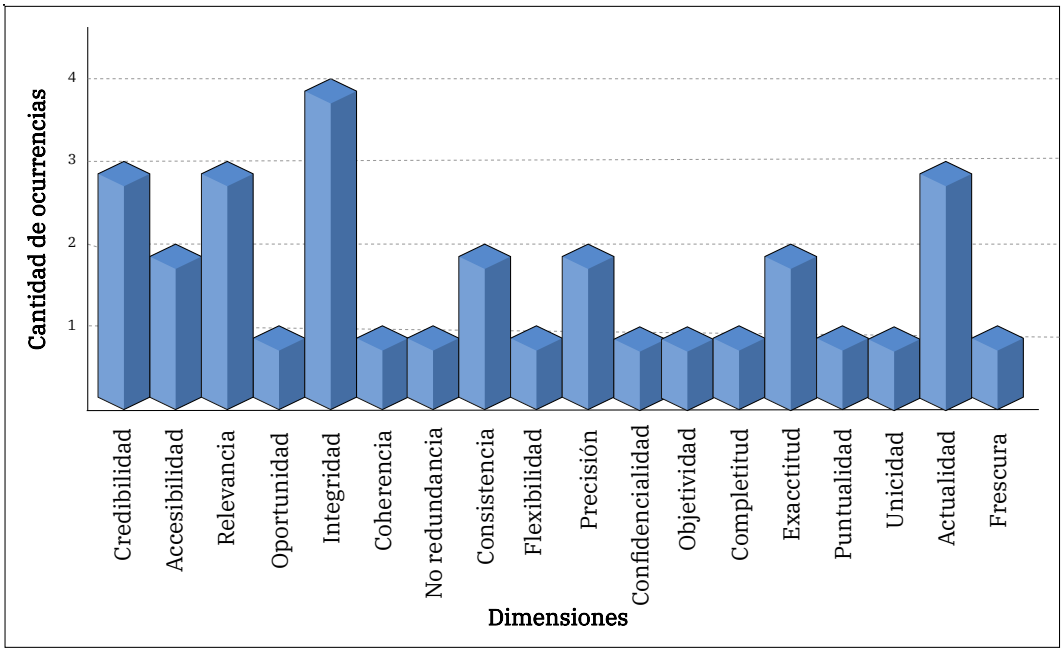


Figura 3.14: Cantidad de ocurrencias de las dimensiones de calidad de datos en las metodologías y frameworks analizados.

3.4.3. Equipo responsable de la gestión de calidad de datos

En la gestión de la calidad de datos propuesta por los frameworks y metodologías analizados se definen actividades, procesos o el uso de herramientas. Para obtener los resultados deseados en la aplicación de un framework o metodología es fundamental garantizar su correcta ejecución. De allí surge la necesidad de contar con un equipo responsable que se encargue de esta tarea.

El framework FW-AGESIC [3] propone la creación de un equipo interdisciplinario compuesto por roles responsables de las diferentes actividades propuestas. Este equipo está compuesto por los siguientes roles:

- Responsable de la aplicación del framework.
- Analista de calidad de datos.
- Técnico de calidad de datos.
- Experto de negocio.
- Experto técnico.

Existen otros roles dentro de la organización que no forman parte del equipo responsable pero su existencia aporta a la gestión de la calidad de datos. En [45] se definen, entre otros, los roles *sponsor* de calidad de datos y usuario de procesos. El rol de *sponsor* de calidad de datos puede ser asignado a una persona o a un grupo que cuente con los recursos económicos y pueda ejercer autoridad. La tarea de este rol consiste en dirigir, financiar y supervisar la gestión de la calidad de datos. Por otro lado, el rol de usuario del proceso incluye a todo empleado de la organización que en sus tareas cotidianas crea o manipula datos. La importancia de reconocer este rol radica en que las diferentes actividades o cambios que se realicen para mejorar la calidad de los datos pueden modificar sustancialmente la forma de realizar sus tareas.

3.4.4. Modelo de madurez

De los frameworks y metodologías analizadas, únicamente el framework DQMP [4] presenta un modelo de madurez. Este modelo se basa en evaluar cinco factores que, en conjunto, definen el grado de madurez de la organización con respecto a la calidad de datos. Los factores a evaluar en el modelo de madurez de DQMP son los siguientes: gestión y planificación, organización y responsabilidades, procesos, conocimientos y competencias y, por último, herramientas. Para cada factor, se presentan varias afirmaciones para evaluar sus valores de verdad.

Este modelo propone cinco niveles de madurez. Cada nivel de madurez define las características mínimas que deben cumplir cada uno de los factores. Por lo tanto, a mayor nivel de madurez, los procesos se encuentran más optimizados y su mejora contribuye a que los datos tengan mejor calidad. Los niveles de madurez propuesto por el framework DQMP son:

- **Reactivo:** en este nivel los procesos de calidad de los datos operan de maneras inesperadas, están mal controlados. Los requisitos de calidad de los datos y el nivel de calidad se desconocen.
- **Controlado:** en este nivel se ha reconocido la necesidad de una gestión de la calidad de los datos. Se han establecido requisitos de calidad de los datos y se han introducido procedimientos que son repetibles.
- **Estandarizado:** en este nivel los procesos de calidad de los datos están estandarizados. La calidad de los datos se verifica para confirmar el cumplimiento de los requisitos propuestos.
- **Administrado:** en este nivel se ha logrado la sostenibilidad de los procesos. Los resultados de las mediciones cualitativas y cuantitativas de la calidad de los datos se utilizan para la gestión de los procesos existentes.

- **Optimizado:** en este nivel los procesos de calidad de los datos se revisan y actualizan periódicamente. Los resultados de las mediciones cualitativas y cuantitativas de la calidad de los datos se utilizan para mejorar los procesos existentes.

3.5. Resumen

En esta sección se presenta un resumen de los aspectos más relevantes presentados en este capítulo. Estos aspectos serán tomados en cuenta en la definición del framework propuesto en el Capítulo 4.

Como se mencionó en la Sección 3.3, en la actualidad no existe un framework de uso genérico para la gestión de la calidad de datos. En consecuencia, surge la necesidad de realizar un estudio detallado de metodologías y frameworks existentes para establecer un punto de partida en la definición de un framework genérico. De los artículos revisados, se seleccionaron seis para ser analizados en profundidad. Cuatro de ellos son metodologías [14][40][39][38] y dos son frameworks [3][4] para la gestión de calidad de datos en diferentes dominios. En la **Figura 3.15** se mencionan estas metodologías y frameworks.

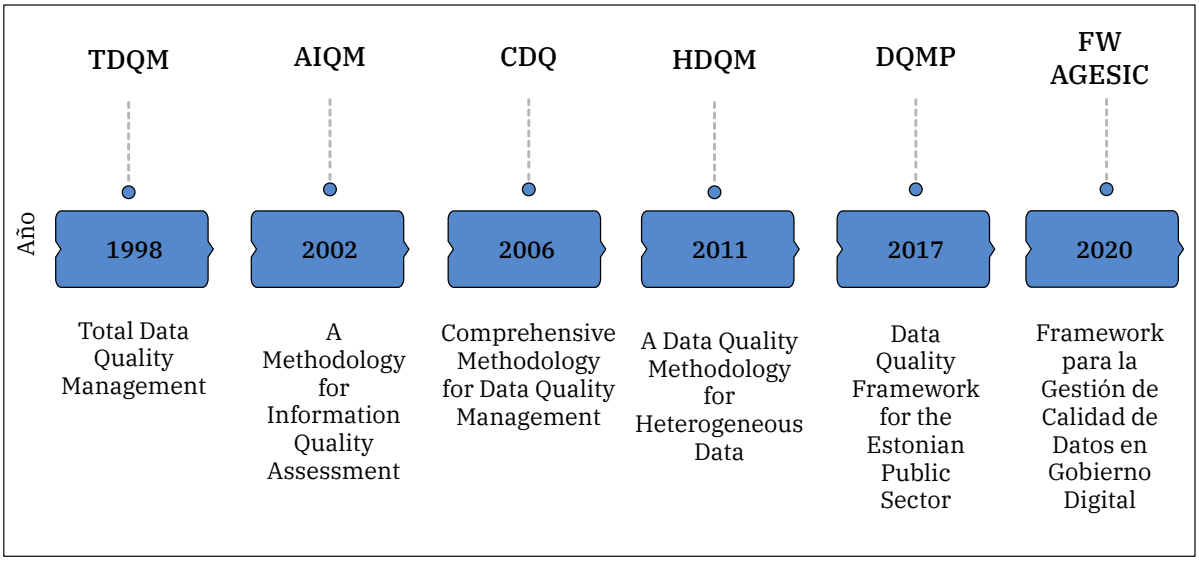


Figura 3.15: Bibliografía analizada, a cada propuesta se le adjunta el año de publicación.

Las metodologías y frameworks tienen características similares. Entre ellas se destacan:

- **Modelo de calidad de datos:** Todos los trabajos analizados incluyen un modelo de calidad de datos. Sin embargo, existen diferentes enfoques. A modo de ejemplo, el framework FW-AGESIC [3] propone un modelo de calidad de referencia a partir del cual se puede crear el modelo que más se adecúe al caso de uso específico. En cambio, el framework DQMP [4] (ambos frameworks se enfocan en el área del sector público correspondiente a su país) propone un modelo de calidad especifico que incluye dimensiones puntuales que satisfacen los requerimientos de calidad del sector público de Estonia.

- **Actividades para la evaluación y mejora en la calidad de datos:** Todas las propuestas analizadas incluyen actividades propias del dominio para la evaluación y mejora de la calidad de datos. En las **Tablas 3.4** y **3.5** se presentan las actividades de cada metodología y framework para las etapas de evaluación y mejora respectivamente.

Por último, es importante destacar otros componentes que no se incluyen en todas las metodologías y frameworks, pero resultaron relevantes durante el análisis. Este es el caso del modelo de madurez, que se incluye únicamente en la metodología DQMP [4]. Otro componente destacable, incluido únicamente en el framework FW_AGESIC [3], es la creación de un equipo de calidad cuya tarea es asegurar la correcta gestión de la calidad de los datos [3].

Actividades para la evaluación de la calidad de datos						
	HDQM	**CDQ**	**AIMQ**	**TDQM**	**FW-AGESIC**	**DQMP**
Clasificación de datos	X					
Medición de la calidad datos	X	X			X	
Reconstrucción del estado		X				
Evaluar datos			X			X
Analizar resultados obtenidos			X			
Desarrollar métricas				X		
Identificar escenario de trabajo					X	X
Definir estrategia de gestión de calidad de datos					X	
Definir modelo de calidad de datos					X	
Definir el alcance de los datos a evaluar						X

Tabla 3.4: Actividades para la etapa de evaluación de calidad de datos.

Actividades para la mejora de la calidad de datos						
	HDQM	**CDQ**	**AIMQ**	**TDQM**	**FW-AGESIC**	**DQMP**
Definición de requisitos de calidad de datos	X	X				
Selección de actividades o técnicas de mejora	X	X		X		X
Elección de proceso de mejora óptimo	X	X				
Elegir mejora para los procesos		X			X	X
Evaluar proceso de mejora		X			X	
Identificar áreas claves para mejorar		X		X		
Definir causas de problemas					X	
Ejecutar plan de mejora					X	

Tabla 3.5: Actividades para la etapa de mejora de calidad de datos presentes en las propuestas analizadas.

Capítulo 4

Propuesta de Framework para la gestión de calidad de datos

Este capítulo tiene como objetivo definir un framework de uso genérico para la gestión de calidad de datos en base a lo presentado en el Capítulo 3. Este framework procura ser de utilidad, independientemente del dominio en el que sea aplicado. En la Sección 4.1 se definen los componentes del framework. Luego, entre las Secciones 4.2 y 4.4 se describen en profundidad cada uno de los componentes presentados.

4.1. Componentes del framework

De acuerdo con la bibliografía consultada, la definición del framework para la gestión de calidad de datos incluye tres componentes, que son los que se presentan en la **Figura 4.1**.

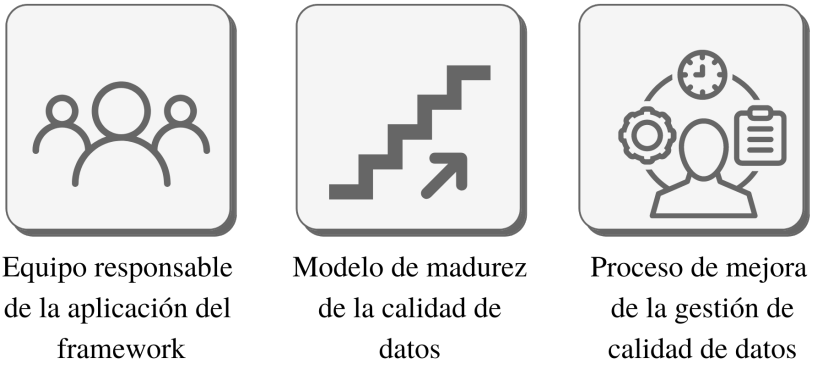


Figura 4.1: Componentes del framework para la gestión de calidad de datos.

Los componentes del framework son: un equipo responsable de la aplicación del framework, un modelo de madurez de calidad de datos y, por último, un proceso de mejora para la misma.

La importancia del **Equipo responsable de la aplicación del framework** radica en formar un grupo de personas responsables del seguimiento y aplicación de las diferentes actividades propuestas en el framework. La inclusión de un **Modelo de madurez de calidad de datos** es relevante debido a que describe un método para conocer el estado actual de los procesos que influyen en la calidad de datos de la organización. Por último, el **Proceso de mejora de calidad de datos** describe las etapas, actividades y roles involucrados en la gestión de la calidad de datos de la organización. Este último es el componente central del framework. En las siguientes secciones se presenta en detalle cada uno de los componentes mencionados.

4.2. Equipo responsable de la aplicación del framework

De acuerdo al análisis realizado en el Capítulo 3, se decide incorporar como uno de los componentes del framework, un equipo interdisciplinario encargado de la gestión del framework. Esta decisión se apoya en el framework propuesto por AGESIC [3] (utilizando los mismos roles), que considera la creación de un equipo responsable encargado de la gestión del framework. Este equipo incluye distintos roles que llevarán a cabo todas las actividades definidas para el desarrollo y gestión del proceso de mejora de calidad de datos. A continuación se describen los roles mencionados.

- **Responsable de calidad de datos:** Es el encargado de la correcta gestión del framework para que se aplique correctamente en el escenario de trabajo. Estará a cargo de documentar todas las actividades realizadas y de asignar las tareas de la aplicación del framework a quien corresponda.
- **Experto de negocio:** Es el referente en el área de negocios. Este rol debe conocer los aspectos fundamentales del negocio para poder evaluar el impacto que tienen los nuevos y actuales requisitos de negocios sobre la calidad de datos. Además, debe ser capaz de documentar dichos requisitos de negocios.
- **Analista de calidad de datos:** Es el experto en el área de calidad y es el responsable del análisis de los aspectos de calidad de los datos más relevantes del escenario definido. Este rol tiene como principal tarea la identificación de los problemas de calidad existentes. Además, debe seleccionar las dimensiones de calidad de datos, los factores de calidad y definir qué métricas de calidad de datos se van a utilizar.
- **Técnico de calidad de datos:** Es el responsable de implementar los recursos técnicos necesarios para la aplicación del framework. Este rol debe implementar, entre otras actividades, los métodos de medición asociados a las métricas definidas.

- **Experto técnico:** Es el experto en los datos que son generados por los sistemas de información incluidos en el escenario de trabajo. Debe tener conocimiento sobre la representación de los datos almacenados.

Existen otros roles que son relevantes, y que no integran el equipo definido anteriormente. Este es el caso del rol **Usuario de procesos**. El mismo comprende a toda persona que, al realizar las actividades definidas para su puesto de trabajo, participa en la creación y/o manipulación de datos. Por lo tanto, su importancia radica en que cualquier cambio que se realice en los procesos de la organización, puede derivar en que se modifique la forma de realizar las tareas asignadas a este rol.

4.3. Modelo de madurez de calidad de datos

En esta sección se define el modelo de madurez de calidad de datos incluido en el framework propuesto. Como se menciona en la sección 2.3.2, un modelo de madurez de calidad de datos tiene como objetivo ayudar a la optimización de la gestión de los datos de una organización. Es una herramienta para que las organizaciones tengan la posibilidad de conocer el estado actual de su gestión de calidad de datos, para luego, generar planes de mejora. En lo que resta de esta sección se definirá el modelo de madurez propiamente dicho. Luego, se explica cómo puede implementarse en una organización.

4.3.1. Definición del modelo de madurez

El modelo de madurez incluido en este framework es propuesto por Angelina Kurniati y Kridanto Surendro [30]. La elección de este modelo de madurez se basa principalmente en la simplicidad de su aplicación. También se considera el hecho de que propone la participación de diferentes roles de la organización para la evaluación de la madurez. Este modelo es una forma de medir qué tan bien desarrollados están los procesos de gestión, es decir, qué tan óptimos son realmente. El modelo consta de seis niveles, los cuales, se utilizan para caracterizar a la organización en relación a un conjunto de atributos de distintas áreas de proceso. Los atributos considerados por los autores de este modelo de madurez son:

- Conocimiento y comunicación (AC, del inglés *Awareness and communication*).
- Políticas, planes y procedimientos (PSP, del inglés *Policies, plans and procedures*).
- Herramientas y automatización (TA, del inglés *Tools and automation*).
- Habilidades y experiencia (SE, del inglés *Skills and expertise*).

- Responsabilidad y rendición de cuentas (RA, del inglés *Responsibility and accountability*).
- Definición y medición de objetivos (GSM, del inglés *Goal setting and measurement*).

Los atributos antes mencionados son evaluados en cada nivel del modelo de madurez, indicando el cambio a implementar para avanzar a un nivel de madurez superior. En consecuencia, a mayor nivel de madurez, los procesos relacionados a los atributos serán más óptimos, brindando mejores resultados. A continuación, se describen los niveles de madurez, indicando el grado de optimización requerido para cada atributo.

Nivel 0: Inexistente
Una organización se encuentra en este nivel cuando no se reconoce ningún tipo de proceso relativo a la gestión de los datos. La organización ni siquiera ha reconocido que es un problema en el que debe enfocarse.

Nivel 1: Inicial
En este nivel, existe evidencia de que la organización reconoce que hay problemas relativos a la gestión de los datos. Sin embargo, no hay procesos estandarizados. Pueden existir enfoques *ad hoc*[1] que se aplican en casos muy básicos o individuales. La gestión se caracteriza por ser desorganizada. A continuación se presentan las características de cada atributo definido para este nivel.

- **AC:** Está surgiendo el reconocimiento de la necesidad en las fases del ciclo de vida de la información. Existe una comunicación esporádica de los temas.
- **PSP:** Existen enfoques *ad hoc* para los procesos y las prácticas en las fases del ciclo de vida de la información. El proceso y las políticas no están definidos.
- **TA:** Algunas herramientas que pueden existir en la fase del ciclo de vida de la información se basan en herramientas de escritorio estándar. No hay un enfoque planificado para el uso de herramientas.
- **SE:** No se identifican las habilidades necesarias para las fases del ciclo de vida de la información. No existe un plan de formación y no se produce una formación formal.
- **GSM:** No existe una definición de rendición de cuentas y responsabilidad en las fases del ciclo de vida de la información. Las personas se apropian de los problemas basándose en su propia iniciativa de forma reactiva.

[1]Expresión que significa literalmente «para esto».

Nivel 2: Repetitivo pero intuitivo

En este nivel los procesos se han desarrollado de modo que diferentes personas que realizan la misma tarea siguen procedimientos similares. No existe capacitación formal o comunicación de procedimientos estándar, y la responsabilidad es de los individuos. Existe un alto grado de confianza en el conocimiento de las personas, lo que puede llegar a generar que se produzcan errores. A continuación, se presentan las características de cada atributo definido para este nivel.

- **AC:** Existe conciencia de la necesidad de actuar en las fases del ciclo de vida de la información. La gerencia comunica los problemas generales.
- **PSP:** En las fases del ciclo de vida de la información surgen procesos similares y comunes, pero son en gran medida intuitivos debido a la experiencia individual. Algunos aspectos del proceso se pueden repetir debido a la experiencia individual, y puede existir cierta documentación y comprensión informal de las políticas y los procedimientos.
- **TA:** Existen enfoques comunes para el uso de las herramientas en las fases del ciclo de vida de la información, pero se basan en soluciones desarrolladas por personas clave. Es posible que las herramientas del proveedor se hayan adquirido, pero probablemente no se hayan aplicado correctamente e incluso pueden ser artículos de estantería.
- **SE:** Se identifican los requisitos mínimos de habilidades para áreas críticas en las fases del ciclo de vida de la información. La capacitación se brinda en respuesta a las necesidades, en lugar de sobre la base de un plan acordado, y se brinda capacitación informal en el trabajo.
- **RA:** Un individuo asume su responsabilidad en las fases del ciclo de vida de la información y generalmente se hace responsable, incluso si esto no se acuerda de manera formal. Hay confusión sobre la responsabilidad cuando surgen problemas y tiende a existir una cultura de culpa.
- **GSM:** Se establecen algunos objetivos en las fases del ciclo de vida de la información. Además, se definen algunas medidas financieras, pero solo la alta dirección las conoce. Existe un monitoreo inconsistente en áreas aisladas.

Nivel 3: Definido

En este nivel los procedimientos se han estandarizado, documentado y son comunicados a través de capacitaciones. Se exige como política de la organización que se sigan estos procesos. Sin embargo, es poco probable que se detecten desviaciones. Los procedimientos no son sofisticados, pero son la formalización de prácticas existentes. A continuación, se presentan las características de cada atributo definido para este nivel.

- **AC:** Se comprende la necesidad de actuar en las fases del ciclo de vida de la información. La gestión es más formal y estructurada en su comunicación.
- **PSP:** Surge el uso de buenas prácticas. El proceso, las políticas y los procedimientos para las fases del ciclo de vida de la información están definidos y documentados para todas las actividades clave.
- **TA:** Se ha definido un plan de uso y estandarización de herramientas para automatizar las fases del ciclo de vida de la información. Las herramientas se utilizan para sus fines básicos, pero es posible que no todas estén de acuerdo con el plan acordado y que no estén integradas entre sí.
- **SE:** Los requisitos de habilidades se definen y documentan para las fases del ciclo de vida de la información en todas las áreas. Se ha desarrollado un plan de formación formal, pero dicha formación se basa en iniciativas individuales.
- **RA:** Se define la responsabilidad y rendición de cuentas de las fases del ciclo de vida de la información. Además, se han identificado los propietarios del proceso. Es poco probable que el propietario del proceso tenga plena autoridad para ejercer las responsabilidades.
- **GSM:** Algunos objetivos y medidas de eficacia se establecen para las fases del ciclo de vida de la información, pero no se comunican y existe un vínculo claro con los objetivos comerciales. Los procesos de medición surgen, pero no se aplican de manera uniforme.

Nivel 4: Gestionado y medible

En este nivel, se monitorea y se mide el cumplimiento de los procedimientos. También, se toman medidas cuando los procesos parecen no estar funcionando de manera efectiva. Los procesos están en constante mejora y proporcionan buenas prácticas. La automatización y las herramientas se utilizan de forma limitada. A continuación, se presentan las características de cada atributo definido para este nivel.

- **AC:** Se comprenden todos los requisitos de las fases del ciclo de vida de la información. Se aplican técnicas de comunicación maduras y se utilizan herramientas de comunicación estándar.
- **PSP:** Las fases del ciclo de vida de la información son sólidas y completas. Se aplican las mejores prácticas internas. Todos los aspectos de las fases del ciclo de vida de la información están documentados y son repetibles.
- **TA:** Las herramientas se implementan en las fases del ciclo de vida de la información de acuerdo con un plan estandarizado y algunas se han integrado con otras herramientas relacionadas. Se están utilizando herramientas en áreas principales para automatizar la gestión de los procesos y monitorear las actividades y controles críticos.
- **SE:** Los requisitos de habilidades en las fases del ciclo de vida de la información se actualizan de forma rutinaria para todas las áreas. Además, se garantiza la competencia en todas las áreas críticas y se fomenta la certificación. Se aplican técnicas de formación maduras de acuerdo con el plan de formación y se fomenta el intercambio de conocimientos. Todos los expertos del dominio interno están involucrados y se evalúa la efectividad del plan de capacitación.
- **RA:** La responsabilidad de las fases del ciclo de vida de la información y la rendición de cuentas se acepta y funcionan, de forma tal, que permite al propietario del proceso cumplir con sus responsabilidades. Existe una cultura de recompensa que motiva la acción positiva.
- **GSM:** La eficiencia y eficacia en las fases del ciclo de vida de la información se miden, comunican y se vinculan con los objetivos comerciales y el plan estratégico de TI. En algunas áreas se utiliza un sistema de gestión que aclara la estrategia y la visión de la organización, traduciéndolas en acciones que se pueden rastrear. Surge una mejora continua.

Nivel 5: Optimizado

Los procesos se han refinado a un nivel donde existen buenas prácticas, y estas se basan en los resultados de la mejora continua y el modelado de madurez con otras organizaciones. Las tecnologías de información se utilizan de forma integrada para automatizar los procesos y flujos de trabajo. Se proporcionan herramientas para mejorar la calidad y la eficacia, haciendo que la organización se adapte rápidamente. A continuación, se presentan las características de cada atributo definido para este nivel.

- **AC:** Existe una comprensión avanzada y prospectiva de los requisitos en las fases del ciclo de vida de la información. Además, se consta de una comunicación proactiva de problemas basados en tendencias. Se aplican técnicas de comunicación maduras y se utilizan herramientas de comunicación integradas.

- **PSP:** Las mejores prácticas y estándares externos se aplican en las fases del ciclo de vida de la información. La documentación del proceso se convierte en flujos de trabajo automatizados. Los procesos, políticas y los procedimientos están estandarizados e integrados para permitir la gestión y la mejora de un extremo a otro.

- **TA:** Los conjuntos de herramientas estandarizados se utilizan en toda la empresa en las fases del ciclo de vida de la información. Las herramientas están completamente integradas con otras herramientas relacionadas para permitir el soporte de los procesos de un extremo a otro. Se están utilizando herramientas para respaldar la mejora del proceso y detectar automáticamente las excepciones de control.

- **SE:** La organización fomenta formalmente la mejora continua de habilidades, basada en objetivos personales y organizacionales definidos en las fases del ciclo de vida de la información. La formación y la educación apoyan las mejores prácticas externas y el uso de conceptos y técnicas de vanguardia. El intercambio de conocimientos es una cultura empresarial y se implementan sistemas basados en el conocimiento. Se utilizan expertos externos y líderes de la industria como orientación.

- **RA:** Los propietarios de procesos tienen la capacidad de tomar decisiones y actuar en las fases del ciclo de vida de la información. La aceptación de la responsabilidad se ha transmitido en cascada a toda la organización de manera constante.

- **GSM:** Existe un sistema integrado de medición del desempeño que vincula el desempeño de TI con los objetivos comerciales mediante la aplicación global de un sistema de gestión en las fases del ciclo de vida de la información. Las excepciones son globales y consistentes por la administración. La mejora continua es una forma de vida.

4.3.2. Implementación del modelo de madurez

La implementación del modelo de madurez consta de tres etapas que pueden observarse en la **Figura 4.2**.

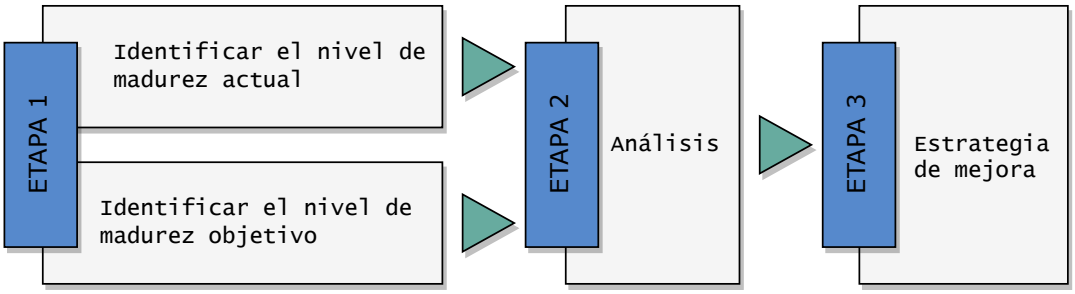


Figura 4.2: Etapas de la implementación del modelo de madurez.

La primera etapa se compone de dos actividades y tiene como objetivo identificar, por un lado, el nivel de madurez actual y por otro, el nivel de madurez objetivo de la organización. Para identificar el nivel de madurez actual de la organización se propone realizar una encuesta a los diferentes roles de la organización que estén involucrados en las distintas etapas del ciclo de vida de la información. Para seleccionar los roles y/o personas que realizarán la encuesta, los autores del modelo de madurez proponen la creación de una matriz R.A.C.I. (del inglés *Responsible, Accountable, Consulted, Informed*). En dicha matriz se mapea cada rol de la organización con cada etapa del ciclo de vida de la información, indicando la responsabilidad o conocimiento que tenga cada rol sobre cada etapa. Esta matriz debe ser utilizada para decidir a quienes consultar. En la **Tabla 4.1** se puede ver un ejemplo de la matriz R.A.C.I. propuesta por este modelo de madurez. Las etapas del ciclo de vida de la información y los roles pueden ser adaptados a la realidad de cada organización. Una vez definidos los roles a los cuales consultar, se debe realizar una encuesta para identificar el nivel de madurez actual y definir el nivel objetivo, basados en el modelo de madurez definido y las diferentes áreas de proceso incluidas.

Fases/Roles	Rol 1	Rol 2	Rol 3	Rol 4
Creación	R	A	C	I
Procesamiento	R	C	A	I
Almacenamiento	A	C		I
Distribución				A

R: Responsible.
A: Accountable.
C: Consulted.
I: Informed.

Tabla 4.1: Ejemplo de matriz R.A.C.I., propuesta por el modelo de madurez, para definir qué roles de la organización deben contestar la encuesta.

La segunda etapa consiste en realizar un análisis de las respuestas obtenidas en la encuesta. Para esto, el modelo de madurez propone crear un diagrama de tipo radar, como el de la **Figura 4.3**, de forma tal de facilitar el análisis de la distancia entre el nivel de madurez actual y objetivo. También puede ser utilizado como guía para implementar un plan de mejora.

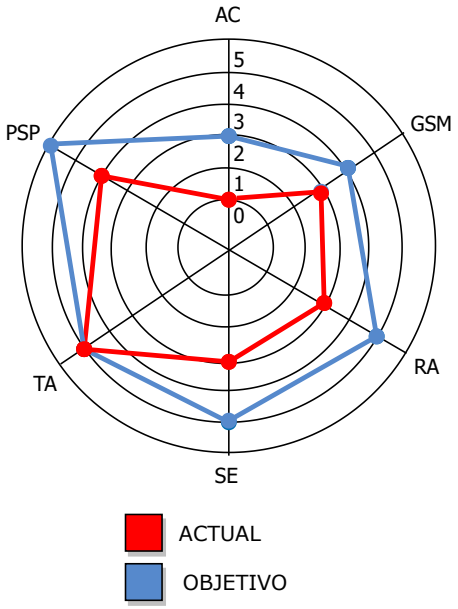


Figura 4.3: Ejemplo de gráfica de tipo radar, propuesta por el modelo de madurez, utilizada para analizar las respuestas a la encuesta.

Por último, en base al resultado del análisis de la encuesta, la organización puede definir un plan de mejora. Generalmente, se priorizan aquellas áreas donde la distancia entre el nivel actual y el objetivo es mayor.

4.4. Proceso de mejora de calidad de datos

En esta sección se define el tercer componente propuesto en el framework. El proceso de mejora de calidad de datos, es un proceso basado en la ejecución sucesiva de etapas, que está diseñado para gestionar la calidad de los datos con el objetivo de mejorarla. Cada etapa del proceso tiene un objetivo bien definido y está compuesto por diferentes actividades. Durante la gestión de cada actividad se generan diferentes productos (documentos, tablas, diagramas, entre otros) que son utilizados en actividades posteriores del proceso.

De acuerdo con lo mencionado en el capítulo anterior, Batini *et al.* [16] identifica dos etapas fundamentales que son comunes a los diferentes enfoques utilizados para la gestión de la calidad

de datos. Dichas etapas son: Evaluación (*Assessment*) y Mejora (*Improvement*). Este framework adopta dichas etapas, sin embargo, a diferencia de lo propuesto por Batini *et al.* [16], en este proceso se propone dividir algunas de las actividades de la Evaluación en cuatro etapas. Dichas etapas son: **Definición del dominio**, **Análisis**, **Definición del modelo de calidad de datos** y **Evaluación de la calidad de datos**. Un motivo de esta división es que se desea que las actividades de cada etapa estén fuertemente relacionadas entre sí, logrando un mejor entendimiento del objetivo de cada etapa. A su vez, como se verá más adelante, en algunos casos se propone re-ejecutar una etapa en función de condiciones definidas en el proceso. La división de etapas propuesta implica que la re-ejecución de una etapa no requiera realizar nuevamente actividades que no sean necesarias. En la **Figura 4.4** se presenta un diagrama que resume las etapas del proceso.

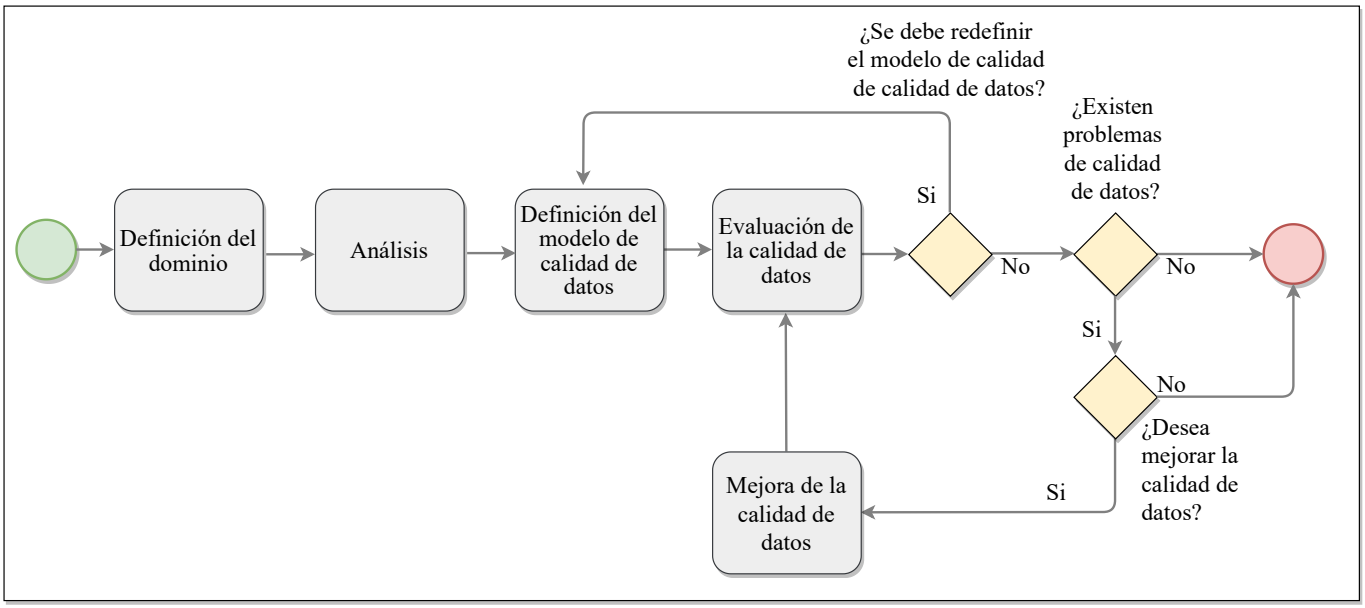


Figura 4.4: Etapas propuestas para el proceso de mejora de calidad de datos.

Las etapas del proceso se explican en detalle entre las Secciones 4.4.1 y 4.4.6. Cada etapa se compone de varias actividades. Para cada actividad, se presenta su objetivo, las entradas necesarias para su gestión, las salidas generadas y los roles responsable de su gestión.

4.4.1. Etapa 1: Definición del dominio

La definición del dominio es la primera etapa del proceso y se encarga de recoger información sobre la organización. Tiene como objetivo identificar el dominio de aplicación del framework. Esto incluye identificar reglas y procesos de negocio, fuentes de datos, políticas de la organización, leyes y los tipos de usuarios de los sistemas involucrados, así como también su vinculación con los procesos de negocio. Además, se busca identificar las organizaciones con las que se intercambian datos, en caso de que existan. Las salidas esperadas para esta etapa son:

- Modelo de procesos que representa a la organización.
- Modelo de datos.
- Documento de reglas, leyes y restricciones de la organización.
- Documento indicando los tipos de usuarios relevados.

Esta etapa surge de la necesidad de realizar una reconstrucción del estado general de la organización, tal como propone Batini *et al.* en la actividad de Análisis para la etapa de Evaluación [16]. Esto también se puede observar en la inclusión del componente Reconstrucción del estado en las metodologías HDQM [38] y CDQ [39] y la etapa Caracterizar el escenario propuesto por el framework de FW_AGESIC [3].

La **Figura 4.5** presenta un diagrama de las actividades definidas en esta etapa. Las mismas se detallan a continuación.

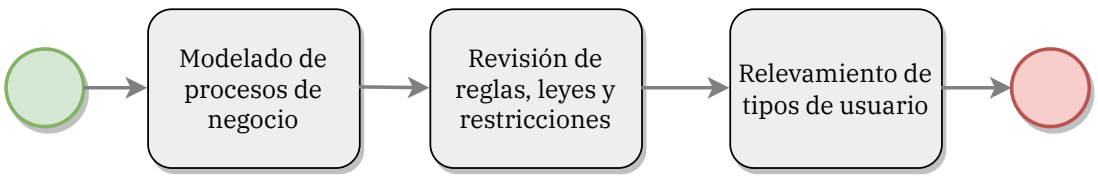


Figura 4.5: Diagrama de las actividades definidas para la etapa Definición del dominio.

Modelado de procesos de negocio y fuentes de datos

Esta actividad consiste en identificar los procesos de negocio de la organización que participan en la creación o manipulación de datos. Además, se identifican las organizaciones externas con las que se intercambian datos y cómo estas influyen en los procesos de negocio. A su vez, se debe representar la estructura y relaciones de los datos objetivo. En la **Figura 4.6** se presenta un ejemplo gráfico reducido, extraído de [3], representando un escenario puntual.

Objetivo: Modelar los procesos, manuales y automáticos, involucrados en la creación, modificación, actualización y transmisión de datos. Se debe tener en cuenta el ciclo de vida de la información y los diferentes sistemas internos y externos involucrados. Modelar las fuentes de datos que se desean analizar.
Responsable: Experto de negocio.
Entrada: Sin entradas.
Salida: Modelo de procesos de la organización y modelo de fuentes de datos. El modelo de procesos de la organización consiste en una representación gráfica y una descripción detallada de cada uno de los procesos de la organización, las entidades involucradas, así como también los tipos de usuarios que actúan. Por otro lado, se deben incluir las relaciones entre las organizaciones externas con las que se intercambian datos y como se comunican entre sí. El modelo de fuentes de datos, consiste en un esquema que permite visualizar la estructura de los datos objetivo y las relaciones que existan entre ellos.

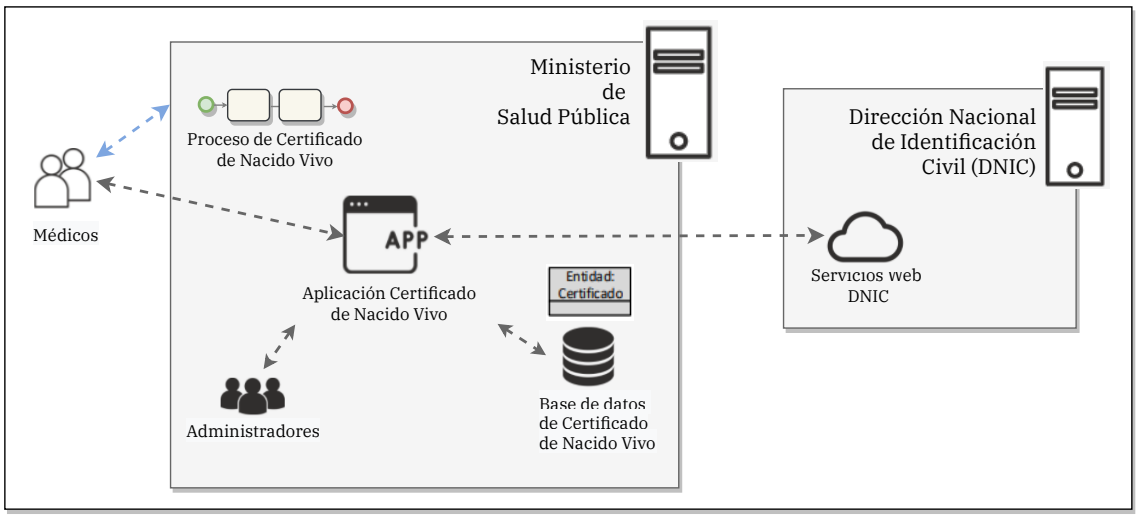


Figura 4.6: Ejemplo gráfico reducido del escenario que enmarca un proceso de negocios. Extraído y adaptado de [3].

Revisión de reglas, leyes y restricciones

Esta actividad consiste en identificar las leyes o restricciones que impactan en los datos o en requisitos sobre su calidad. En la **Tabla 4.2** se presenta un ejemplo del registro de las reglas, leyes y/o restricciones.

Objetivo: Identificar toda normativa, ley o restricción, tanto interna como externa a la organización, que pueda derivar en un requerimiento de calidad de datos o impactar directamente en los datos.
Responsable: Experto de negocio.
Entrada: Sin entradas.
Salida: Documento de reglas y restricciones. Debe incluir un identificador y descripción para cada restricción identificada.

Identificador	**Descripción**
#1	Ley N.º 18331: Ley de protección de datos personales.
...	...

Tabla 4.2: Ejemplo reducido del documento para el registro de reglas, leyes y restricciones.

Relevamiento de tipos de usuario

En esta actividad se identifican, dentro de la organización, los roles de usuarios que manipulan datos de forma directa e indirecta. En la **Tabla 4.3** se presenta un ejemplo reducido de la plantilla a utilizar para el registro de los roles de usuarios involucrados en los procesos de negocio identificados.

Objetivo: Relevar los diferentes tipos de usuarios que participan en las etapas del ciclo de vida de la información.
Responsable: Experto de negocio.
Entrada:: Sin entradas.
Salida: Documento de tipos de usuarios. En este documento se deben incluir todos los roles de usuarios que participen en la creación y manipulación de datos e información, así como también la identificación de los procesos con los que interactúan.

Identificador	**Nombre**	**Descripción**	**Procesos con los que interactúa**
#1	Administrador	Rol encargado de la administración de la aplicación.	Aplicación en general.
...	...	...	...

Tabla 4.3: Ejemplo reducido del relevamiento de roles de usuarios.

4.4.2. Etapa 2: Análisis

La etapa de análisis es la segunda etapa del proceso y es la encargada del relevamiento de los requisitos y problemas de calidad de datos existentes en la organización. Esta etapa tiene como objetivo obtener una primera aproximación al conocimiento de los datos e investigar los problemas de calidad de datos de la organización. Esto es de suma importancia ya que influye en la definición de la estrategia para la gestión de calidad de datos, así como también en la definición del plan de mejora. Las salidas esperadas para esta etapa son:

- Informe de *Data Profiling*.
- Documento de problemas de calidad de datos.
- Documento de requisitos de calidad de datos.

La **Figura 4.7** presenta un diagrama de las actividades incluidas en esta etapa. La inclusión de estas actividades se apoya en las actividades de la etapa de Evaluación propuesta por Batini *et al.* [16]. Estas actividades también se incluyen en la metodología TDQM [14] y el framework FW_AGESIC [3].

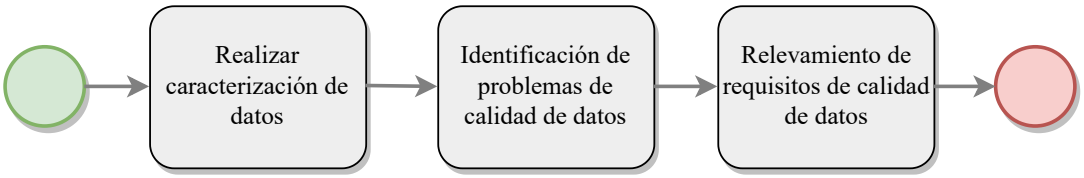


Figura 4.7: Diagrama de flujo de actividades de la etapa Análisis.

Realizar caracterización de datos

En esta actividad se realiza una introducción a las características más destacables de los datos. Es decir, consiste en analizar un conjunto de datos, y a partir de estos identificar sus propiedades y características. Es una actividad fuertemente centrada en el análisis de los atributos, incluyendo su contenido, estructura, relaciones de dependencia, valores faltantes o duplicados, entre otros. En la **Tabla 4.4** se presenta un ejemplo del informe de *data profiling* reducido a un único atributo. El informe de *data profiling* presentado en el ejemplo no impide agregar otras propiedades o características relevantes para cada dominio.

Objetivo: Analizar un conjunto de datos identificando sus propiedades y características.
Responsable: Analista de calidad de datos y experto técnico.
Entrada: Fuentes de datos incluidas en el dominio de aplicación.
Salida: Informe de *Data profiling*. En este informe se detallan las características y propiedades más importantes de los datos.

Propiedad	**Valor**
Atributo	Primer apellido
Tabla en BD	PERSONA
Nombre de atributo en BD	Apellido
Tipo de atributo en BD	VARCHAR(100 Byte)
Restricciones sobre atributo en BD	No nulo
Cantidad total de registros	150.000
Valores duplicados	1.000
Valores nulos	100

Tabla 4.4: Ejemplo reducido de caracterización de una variable.

Identificación de problemas de calidad de datos

En esta actividad, se busca reconocer y clasificar los problemas de calidad de datos. Estos problemas deben clasificarse de acuerdo a su relevancia y/o gravedad. Para simplificar, se utilizaran valores de gravedad alta, media o baja. Sin embargo, también pueden utilizarse otros criterios para definir la gravedad. En la **Tabla 4.5** se presenta un ejemplo reducido del documento de problemas de calidad.

Objetivo: Identificar y clasificar los problemas de calidad de datos.
Responsable: Analista de calidad de datos y responsable de calidad de datos.
Entrada: Modelo de procesos de la organización e informe de *data profiling*.
Salida: Documento de problemas de calidad de datos. Este documento incluye un listado de los problemas de calidad identificados en la organización y clasificados según su gravedad.

Identificador	Descripción	Gravedad
#P1	**Correo electrónico inexistente:** Un problema común es que a muchos usuarios no les llegan los e-mails enviados por el sistema porque ingresaron incorrectamente su dirección de correo.	Media
#P2	**Fecha de nacimiento poco confiable:** Se han detectado fechas que no se encuentran en el rango de edades de usuarios del sistema (18 y 55 años).	Alta

Tabla 4.5: Ejemplo reducido del informe de problemas de calidad de datos identificados.

Relevamiento de requisitos de calidad de datos

En esta actividad se realiza el análisis e identificación de los requisitos de calidad de datos existentes que son relevantes para la organización. Para realizar esta actividad, deben utilizarse las salidas producidas en la etapa anterior. En la **Tabla 4.6** se presenta un ejemplo reducido de la planilla a utilizar para la identificación de los requisitos de calidad de datos y su priorización.

Objetivo: Identificar, analizar y priorizar los requisitos de calidad de datos de la organización.
Responsable: Responsable de calidad de datos y analista de calidad de datos.
Entrada: Documento de reglas, leyes y restricciones, modelo de procesos de la organización, documento de problemas de calidad de datos e informe de *data Profilling*.
Salida: Documento de requisitos de calidad de datos. Este documento debe incluir un listado priorizado con el detalle de los requisitos de calidad de datos.

Identificador	Descripción	Prioridad
#REQ1	El e-mail que registra el usuario debe tener un formato de correo válido.	Alta
#REQ2	La fecha de nacimiento que registra el usuario debe encontrarse entre el rango de edades permitidas del sistema.	Media

Tabla 4.6: Ejemplo reducido de los requisitos de calidad de datos.

4.4.3. Etapa 3: Definición del modelo de calidad de datos

La definición del modelo de calidad de datos es la tercera etapa del proceso. El principal objetivo es definir el modelo de calidad de datos. Para definir el modelo se deberán contemplar los documentos e informes generados en la etapa anterior. Por otro lado, en esta etapa se desarrolla el software que implementa las métricas definidas en el modelo de calidad de datos. Por lo tanto, luego de realizadas las actividades de esta etapa, se obtendrán las siguientes salidas:

- Modelo de calidad de datos.
- Software que implementa las métricas definidas.

La **Figura 4.8** presenta un diagrama para las actividades definidas en esta etapa. La inclusión de estas etapas se apoya en la etapa de Evaluación de Batini *et al.* [16]. Estas actividades también se incluyen en todas las metodologías y frameworks analizados en el Capítulo 3 [38][39][40][14][3][4].

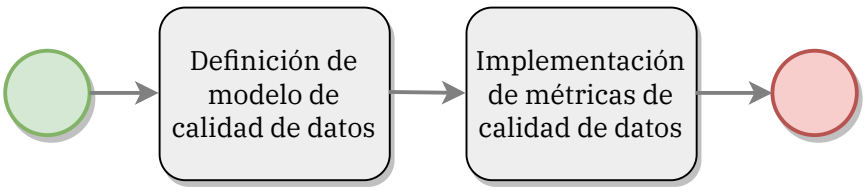


Figura 4.8: Diagrama de flujo de actividades de la etapa Definición del modelo de calidad de datos.

Definición de modelo de calidad de datos

En esta actividad se define el modelo de calidad de datos, es decir las dimensiones, factores, métricas y métodos de medición. La definición del modelo de calidad de datos debe orientarse a resolver los problemas y requisitos de calidad de la organización. En la **Tabla 4.7** se presenta un ejemplo reducido de un modelo de calidad de datos para una variable en particular.

Objetivo: Definir las dimensiones de calidad de datos, sobre qué datos se aplican y cómo se miden. Especificar factores, métricas y métodos de medición para cada dimensión considerada.
Responsable: Analista de calidad de datos y responsable de calidad de datos.
Entrada: Documento de requisitos de calidad de datos, documento de problemas de calidad de datos.
Salida: Modelo de calidad de datos.

Dimensión	**Factor**	**Métrica**	**Procedimiento de medición**
Unicidad	Unicidad atributo	Único (Correo electrónico)	Consulta SQL buscando otra tupla que contenga el mismo valor para ese atributo.
Completitud	Densidad	Nulos (Correo electrónico)	Comparar con NULL.
...	...	...	...

Tabla 4.7: Ejemplo reducido del modelo de calidad de datos para una variable.

Implementación de métricas de calidad de datos

Esta actividad consiste en implementar las herramientas de software necesarias para realizar las mediciones. No existen restricciones en cuanto a la tecnología utilizada para el desarrollo de software.

Objetivo: Implementar los métodos de medición definidos para las métricas incluidas en el modelo de calidad de datos.
Responsable: Técnico de calidad de datos.
Entrada: Modelo de calidad de datos.
Salida: Software que implementa las métricas definidas.

4.4.4. Etapa 4: Evaluación de la calidad de datos

La cuarta etapa del proceso es la evaluación de la calidad de datos. El principal objetivo en esta etapa es determinar el nivel de calidad de datos de acuerdo al modelo definido anteriormente. También se propone realizar una reevaluación de los problemas de calidad de datos existentes. Esto se debe a que luego de medir la calidad de datos pueden identificarse nuevos problemas. En esta etapa se generan las siguientes salidas:

- Informe de medición de calidad de datos.
- Documento de problemas de calidad de datos refinado.

La **Figura 4.9** presenta un diagrama del flujo de las actividades definidas para esta etapa. La inclusión de estas etapas se apoya en las actividades de la etapa de Evaluación y Mejora de Batini *et al.* [16]. Estas actividades también se incluyen en las metodologías y frameworks analizados en el Capítulo 3 [38][39][40][14].

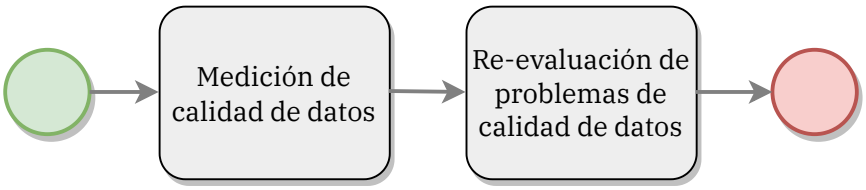


Figura 4.9: Diagrama de flujo de actividades de la etapa Evaluación de calidad de datos.

Medición de calidad de datos

En esta actividad se debe realizar la medición de la calidad de los datos objetivo. En la **Tabla 4.8** se presenta un ejemplo reducido del informe con el desglose de las mediciones de calidad de datos realizadas.

Objetivo: Realizar las mediciones ejecutando los métodos de medición implementados.
Responsable: Técnico de calidad de datos.
Entrada: Modelo de calidad de datos.
Salida: Informe de medición de calidad de datos. Se debe registrar el resultado obtenido para cada métrica definida.

Métrica	Resultado	Descripción
Densidad de dirección de correo electrónico	93,5 %	La densidad es alta. Existe un 6.5 % de valores nulos.

Tabla 4.8: Ejemplo de informe de medición de calidad de datos reducido a una variable.

Re-evaluación de problemas de calidad de datos

En esta actividad, se realiza una revisión de los problemas de calidad identificados hasta el momento. El objetivo consiste en identificar los problemas de calidad de datos que no fueron detectados en etapas anteriores e incluirlas en el documento definido anteriormente.

Objetivo: Re-evaluar los problemas de calidad existentes. Se busca identificar nuevos problemas de calidad de datos.
Responsable: Responsable de calidad de datos.
Entrada: Informe de medición de calidad de datos y documento de problemas de calidad de datos.
Salida: Documento actualizado de problemas de calidad de datos.

Una vez finalizada la etapa **Evaluación de la calidad de datos** el proceso propone la evaluación de condiciones para determinar el flujo de ejecución que se debe seguir (ver **Figura 4.4**). En primer lugar, se debe evaluar si se requieren cambios en el modelo de calidad de datos utilizado para la evaluación. En caso afirmativo, se debe volver a la etapa **Definición del modelo de calidad de datos** y re-ejecutar sus actividades. Por el contrario, si no se desea modificar el modelo de calidad, se debe verificar si existen problemas de calidad de datos. En caso de que no existieran, se da por finalizado el proceso, debido a que no sería necesario realizar una etapa de mejora. Por el contrario, si existen problemas de calidad, se brinda la posibilidad de decidir si se quiere realizar mejoras que solucionen los problemas identificados. La decisión de no realizar mejoras implica la finalización del proceso. En caso de que la decisión sea realizar mejoras, el flujo del proceso continúa con la ejecución de la etapa **Mejora de calidad de datos**.

4.4.5. Etapa 5: Mejora de la calidad de datos

La etapa de mejora de la calidad de datos es la última etapa del proceso. El objetivo de esta etapa consiste en el diseño y gestión de estrategias enfocadas en mejorar la calidad de los datos. La estrategia debe enfocarse en resolver los problemas de calidad de datos de la organización y adaptarse a sus requisitos. En esta etapa las salidas esperadas son:

- Plan de acción.
- Informe de ejecución del plan de acción.

La **Figura 4.10** presenta un diagrama del flujo de las actividades para esta etapa. La inclusión de estas etapas se apoya en las actividades de la etapa de Mejora de Batini *et al.* [16]. Estas actividades también se incluyen en algunas metodologías y frameworks analizados en el Capítulo 3.

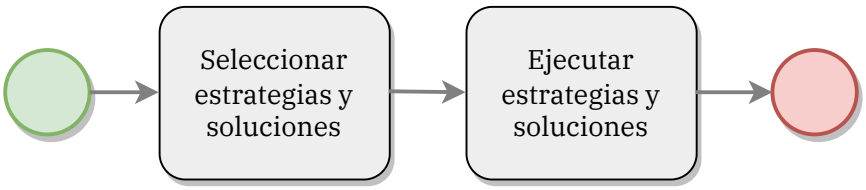


Figura 4.10: Diagrama de flujo de actividades de la etapa Mejora de calidad de datos.

Seleccionar estrategias y soluciones

En esta actividad se define un plan de acción que permite resolver los problemas de calidad de datos identificados. La definición del plan de acción depende de los siguientes aspectos:

- Los problemas de calidad de datos identificados.
- Los objetivos de calidad de datos.
- Modelo de madurez.
- Las tareas que afectan los problemas de calidad.
- Los usuarios de los sistemas que producen datos.

Las acciones definidas pueden incluir la implementación de los cambios necesarios (software o procesos) así como también, técnicas de limpieza de los datos existentes. La salida de esta actividad es el documento de plan de acción. Este consiste en un informe detallado de las soluciones

y tareas a realizar con el objetivo de mejorar la calidad de datos. Para cada tarea definida se debe indicar un responsable. En la **Figura 4.9** se presenta un ejemplo reducido del plan de acción.

Objetivo: Definir las estrategias y técnicas adecuadas a los problemas de calidad a resolver.
Responsable: Responsable de calidad de datos y analista de calidad de datos.
Entrada: Informe de medición de calidad de datos y documento de problemas de calidad refinado.
Salida: Documento de plan de acción.

Identificador	**Estrategia**	**Descripción**	**Problema a resolver**	**Responsable**
#A1	Correctitud de correo electrónico	Se debe controlar en el formulario donde se ingresa el correo electrónico que este tenga el formato correcto.	Problema #P1 **Ref: 4.5**	Técnico de calidad de datos

Tabla 4.9: Ejemplo reducido del plan de acción.

Ejecutar estrategias y soluciones

En esta actividad se ejecutan y documentan los resultados para cada estrategia definida en el plan de acción. Cada estrategia cuenta con un rol responsable de su ejecución. El responsable es quién debe supervisar la correcta implementación de la tarea. En la **Figura 4.10** se presenta un ejemplo reducido del informe del plan de acción.

Objetivo: Ejecutar el plan de acción definido y registrar los resultados obtenidos.
Responsable: Responsable de calidad de datos y técnico de calidad de datos.
Entrada: Plan de acción.
Salida: Informe de ejecución del plan de acción.

Identificador	**Estrategia**	**Resultado**
#A1	Correctitud de correo electrónico	100 % datos correctos

Tabla 4.10: Ejemplo reducido del informe del plan de acción.

4.4.6. Monitoreo de calidad de datos

En las secciones anteriores se presentó en detalle cada una de las etapas y actividades que componen el proceso de mejora de calidad de datos propuesto en este framework. Tal como lo incluye Batini *et al.* en su etapa de Mejora [16], surge la necesidad de realizar actividades de control de calidad de datos posterior a dicha gestión. Esto requiere determinar si las acciones implementadas generan el resultado esperado en el corto, mediano y largo plazo. En esta sección, se introducen las actividades propuestas en este framework para implementar una etapa de monitoreo de calidad de datos.

El proceso de mejora de calidad de datos, propone la ejecución secuencial de etapas cuyo objetivo, a más alto nivel, es mejorar la calidad de los datos. Sin embargo, no se puede asegurar que, luego de ejecutadas las etapas del proceso, la calidad de los datos alcanzada se mantenga en el tiempo. Se identifican dos posibles causas que derivan en la pérdida de calidad luego de completado el proceso de mejora de calidad de datos.

- Los nuevos datos que producen los sistemas de información, de forma constante, pueden disminuir la calidad de datos.
- Los cambios, tanto en los sistemas de información propios, como en factores ajenos pueden introducir nuevos problemas y/o requisitos de calidad de datos.

En las siguientes subsecciones se definen las actividades para implementar el monitoreo de calidad de datos.

Control de calidad de datos

Luego de ejecutado el proceso de mejora, no se puede asegurar que la calidad de los datos se mantenga en los niveles alcanzados anteriormente. Para esto se debe implementar de forma continua un control de la calidad de los datos utilizando el modelo de calidad de datos definido. Este control debe realizarse periódicamente de acuerdo a las necesidades de cada dominio.

En la **Figura 4.11** se observan las actividades que componen el control de calidad de datos. Estas actividades consisten en ejecutar periódicamente las etapas **Evaluación de calidad de datos** y **Mejora de calidad de datos** definidas en el proceso de mejora de calidad de datos.

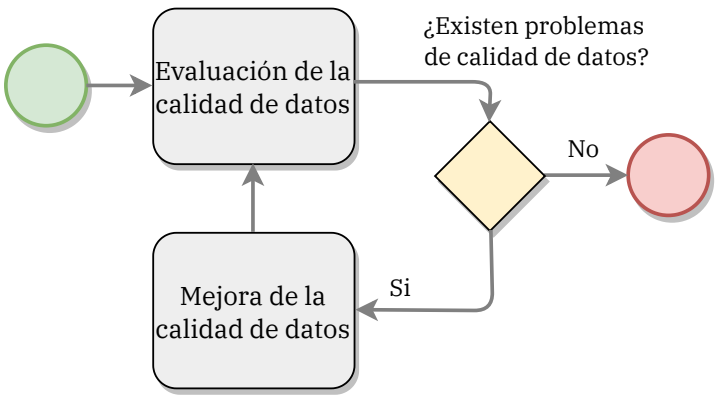


Figura 4.11: Actividades para el control de calidad de datos.

Control de requisitos de calidad de datos

Otra causa que puede disminuir la calidad de los datos es el cambio en los sistemas de información. Se consideran aquellos cambios que pueden agregar o modificar requisitos de calidad de los datos. Algunos ejemplos de los cambios mencionados son, cambios en los requerimientos del sistema, cambios en las estructuras de datos utilizadas por el sistema de información o cambios en sistemas externos de los cuales se obtienen y almacenen datos.

El control de requisitos propuesto en este framework consiste en planificar una nueva ejecución del proceso de mejora definido (ver **Figura 4.4**). La periodicidad con la que se vuelve a ejecutar este proceso debe ser definida teniendo en cuenta las necesidades de cada organización.

Capítulo 5

Implementación del Framework en Eclipse Process Framework Composer

Este capítulo tiene como objetivo brindar la documentación de la implementación del framework de uso genérico para la gestión de calidad de datos definido en el capítulo anterior, utilizando la herramienta *Eclipse Process Framework Composer* (EPF *Composer*). En la Sección 5.1 se presenta la herramienta utilizada para la construcción del framework. Luego, en la Sección 5.2 se expone la implementación de este.

5.1. Eclipse Process Framework Composer

Eclipse Process Framework Composer (de ahora en más EPF *Composer*), es una plataforma de herramientas de código abierto que permite crear, adaptar y publicar distintos métodos y procesos de organizaciones o proyectos individuales [46]. El objetivo de EPF *Composer* radica en la solución de dos problemas claves en la ejecución de un nuevo proceso [47]. El primero de ellos surge de la necesidad de capacitar a los equipos, en los métodos aplicables a las funciones de las que son responsables. El segundo objetivo, surge de la necesidad que representan los equipos al momento de comprender cómo aplicar estos métodos a lo largo del ciclo de vida de desarrollo y la gestión de cambios de la organización.

Para lidiar con estos problemas EPF *Composer* tiene dos propósitos. El primero consiste en brindar una base de conocimientos de capital intelectual que permita explorar, administrar e implementar contenido. Este contenido se puede adaptar a cualquier proceso de una organización y consta de, por ejemplo, definiciones de métodos, documentos técnicos, mejores prácticas, procedimientos y reglamentos internos o material de capacitación, entre otros. El segundo propósito es proveer funcionalidades de ingeniería de procesos para apoyar a los gerentes de proyecto

en la selección, adaptación e implementación de procesos en los proyectos de desarrollo de las organizaciones [47].

En función de lo anterior, en la siguiente sección, se presenta la implementación del proceso de mejora de calidad de datos propuesto en este proyecto incluyendo etapas, actividades, roles y productos de trabajo que componen al framework.

5.2. Implementación

En esta sección se presenta la implementación del framework propuesto en este proyecto usando la herramienta EPF *Composer* presentada en la sección anterior. Antes de comenzar a presentar la implementación del framework haremos una breve mención de la nomenclatura utilizada por la herramienta para definir los componentes del framework. En este informe, se utilizaron los términos **etapa** y **actividad** para definir los componentes del proceso de mejora de calidad de datos propuesto. Mientras que, en EPF *Composer*, para etapa y actividad se utilizan los términos **fase** y **tarea** respectivamente. Por otro lado, los productos de trabajo en EPF Composer se corresponden con las entradas y salidas de las actividades propuestas en el proceso de mejora de calidad de datos. La **Tabla 5.1** muestra el mapeo entre la nomenclatura de EPF *Composer* y el framework propuesto.

Nomenclatura de EPF Composer	**Artefactos del framework**
Tarea	Actividad
Fase	Etapa
Producto de trabajo	Entradas/Salidas de las actividades.

Tabla 5.1: Mapeo EPF Composer y framework propuesto.

EPF *Composer* permite la creación de una biblioteca que almacena un *core* de contenido del método y la definición de procesos. Esta biblioteca permite gestionar el *core* pudiendo crear o editar disciplinas, roles, tareas y productos de trabajo entre otros. Una vez definidos los elementos del *core*, estos pueden utilizarse (y reutilizarse), en la creación de distintos procesos. En la **Figura 5.1** se presenta un ejemplo de la biblioteca mencionada anteriormente.

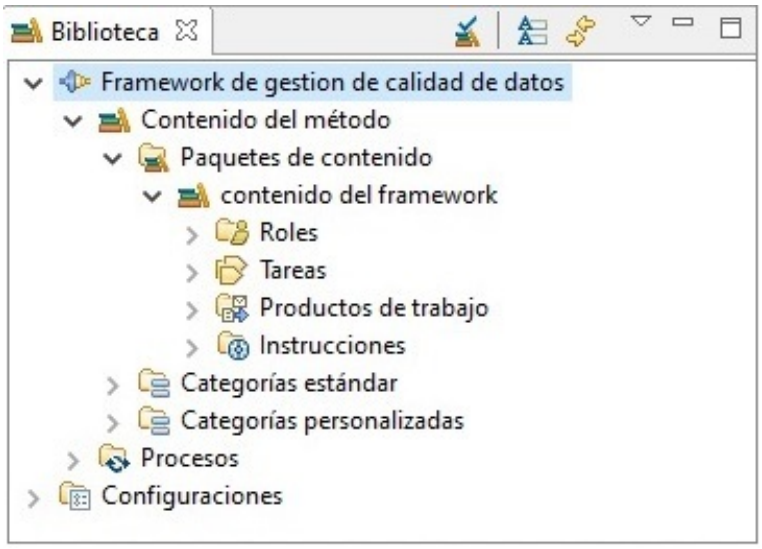


Figura 5.1: Interfaz para la gestión de la biblioteca de los métodos del framework utilizando la herramienta EPF *Composer*.

El primer paso en la implementación fue crear los roles del equipo de calidad de datos. Para esto, la herramienta brinda una interfaz gráfica que permite configurar distintos atributos como el nombre, una descripción breve y otra general, entre otros. En la **Figura 5.2** se muestra un ejemplo de la interfaz para crear y editar los roles.

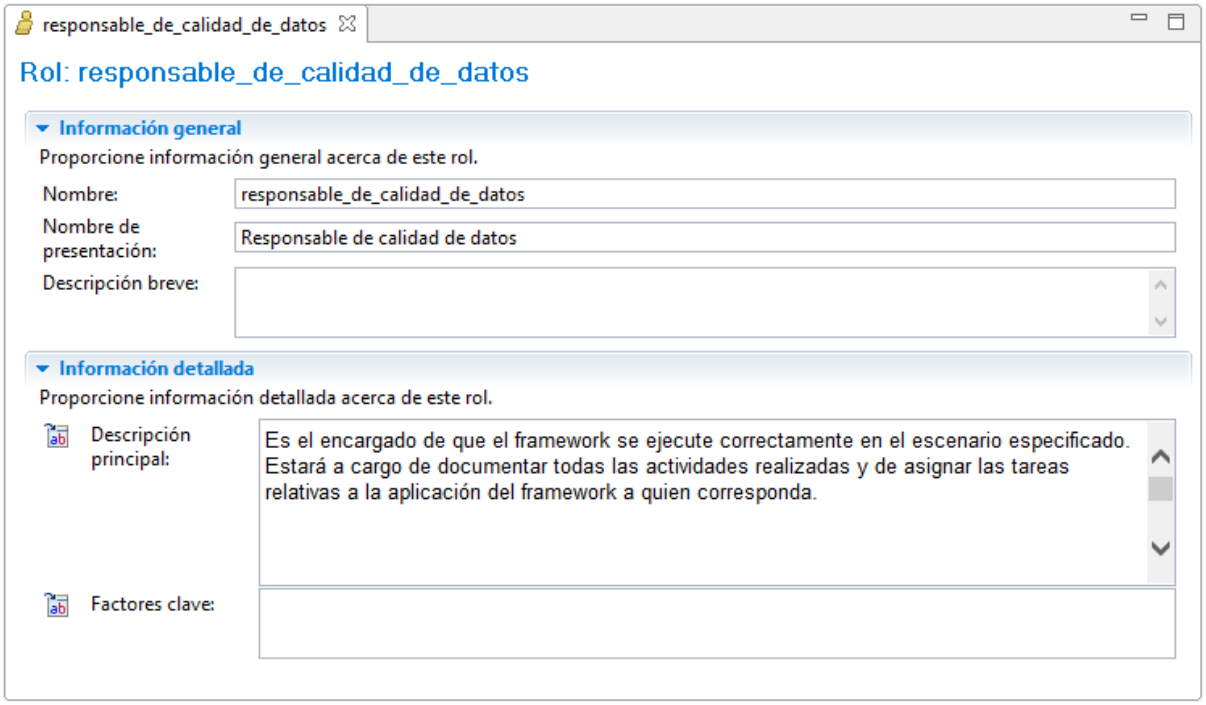


Figura 5.2: Interfaz de creación y edición del rol responsable de calidad de datos utilizando la herramienta EPF *Composer*.

Luego de creados todos los roles propuestos en el framework, se crearon los productos de trabajo utilizados en las tareas. Un producto de trabajo hace referencia tanto a una entrada, como a una salida de una actividad. Cada producto de trabajo incluye una descripción y, para los que aplique, una plantilla con el formato que debe tener el documento de salida. En la figura **Figura 5.3** se presenta un ejemplo que muestra la creación de un producto de trabajo.

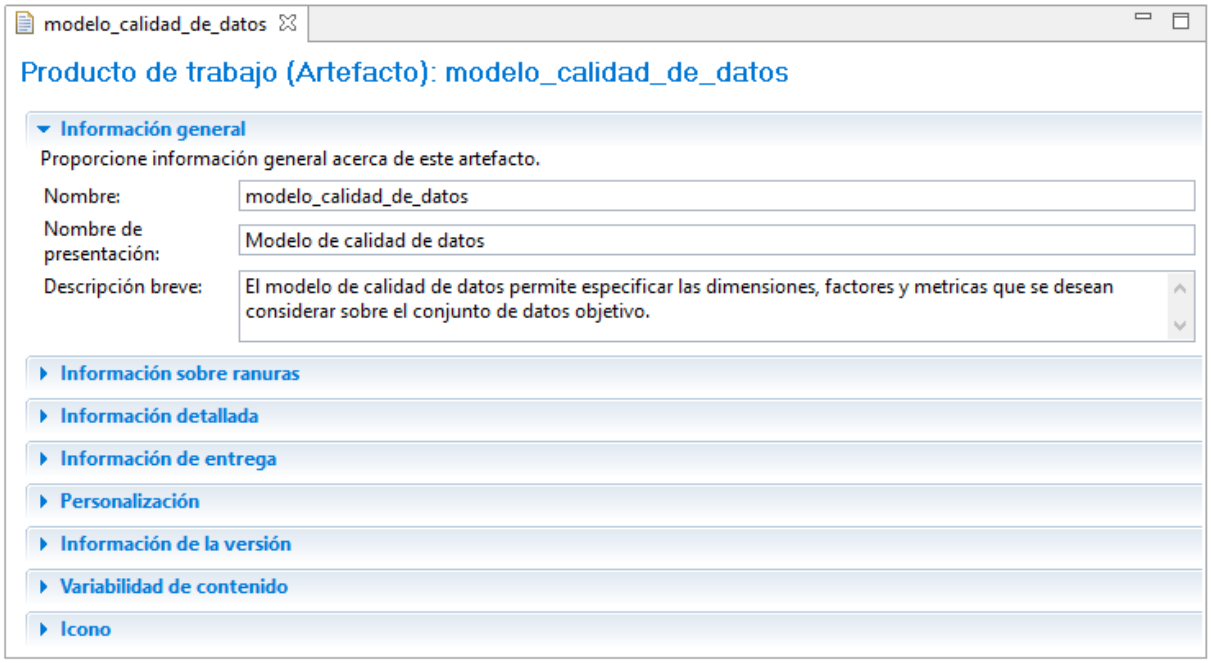


Figura 5.3: Interfaz de creación y edición del producto de trabajo: modelo de calidad de datos.

Una vez creados los roles y los productos de trabajo se crean las tareas. Las tareas, además de contar con un nombre, descripción y objetivos, pueden vincularse con roles y productos de trabajo. Un rol puede ser el realizador principal o el secundario de una tarea. Un producto de trabajo puede asociarse a una tarea como entrada y/o salida. La herramienta permite la creación de disciplinas que agrupan a las tareas de un mismo tipo. En este proyecto se crearon tres disciplinas para agrupar las distintas tareas propuestas. Estas disciplinas son Calidad de datos, Desarrollo y Análisis. En la **Figura 5.4** se presenta la interfaz que permite crear y editar las tareas.

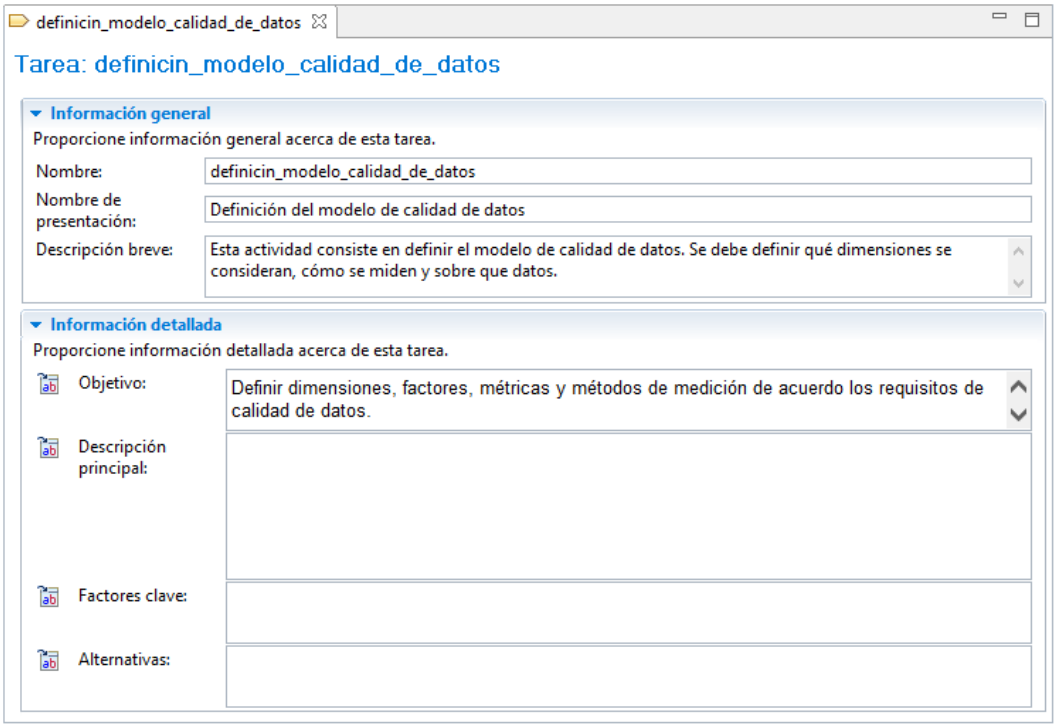
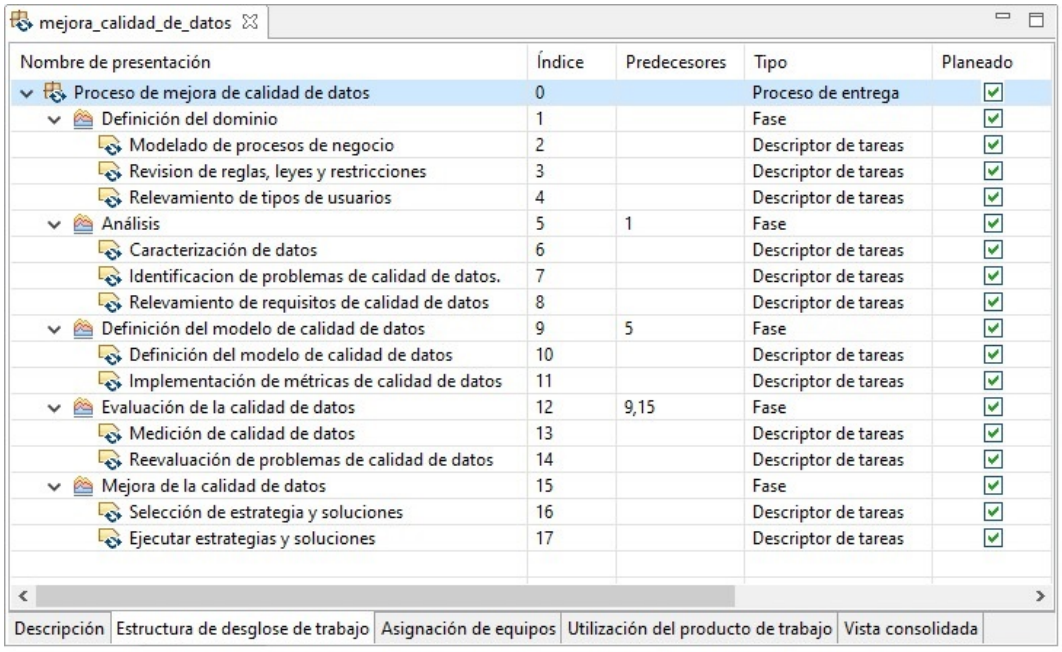


Figura 5.4: Interfaz de creación y edición de tareas.

El último paso en la implementación consiste en la creación de un proceso con cinco fases. Para dichas fases se incluyen las tareas creadas anteriormente. En las tareas agregadas del proceso definido se pueden visualizar los vínculos creados entre las fases, tareas, roles y productos de trabajo. En la **Figura 5.5** se presenta la interfaz que permite crear y editar procesos.



Nombre de presentación	Índice	Predecesores	Tipo	Planeado
Proceso de mejora de calidad de datos	0		Proceso de entrega	☑
Definición del dominio	1		Fase	☑
Modelado de procesos de negocio	2		Descriptor de tareas	☑
Revision de reglas, leyes y restricciones	3		Descriptor de tareas	☑
Relevamiento de tipos de usuarios	4		Descriptor de tareas	☑
Análisis	5	1	Fase	☑
Caracterización de datos	6		Descriptor de tareas	☑
Identificacion de problemas de calidad de datos.	7		Descriptor de tareas	☑
Relevamiento de requisitos de calidad de datos	8		Descriptor de tareas	☑
Definición del modelo de calidad de datos	9	5	Fase	☑
Definición del modelo de calidad de datos	10		Descriptor de tareas	☑
Implementación de métricas de calidad de datos	11		Descriptor de tareas	☑
Evaluación de la calidad de datos	12	9,15	Fase	☑
Medición de calidad de datos	13		Descriptor de tareas	☑
Reevaluación de problemas de calidad de datos	14		Descriptor de tareas	☑
Mejora de la calidad de datos	15		Fase	☑
Selección de estrategia y soluciones	16		Descriptor de tareas	☑
Ejecutar estrategias y soluciones	17		Descriptor de tareas	☑

Figura 5.5: Interfaz de creación y edición de procesos.

Una vez culminada la implementación del framework, la herramienta permite publicar un sitio web donde se puede visualizar y navegar por todos los componentes definidos. En las **Figuras 5.6** y **5.7** se pueden observar algunas de las vistas incluidas en el sitio web publicado.

Figura 5.6: Vista del sitio web para la visualización y navegación del proceso y sus componentes.

Figura 5.7: Vista del sitio web para la visualización y navegación de la etapa de definición del dominio.

Capítulo 6

Aplicación de la propuesta

Este capítulo tiene como objetivo mostrar, a través de un caso de estudio real, la aplicación del proceso de mejora propuesto en el Capítulo 4. En la Sección 6.1 se presenta una descripción general del caso de estudio seleccionado y los problemas de calidad de datos más relevantes, involucrados en dicho caso. En la Sección 6.2 se ejecutan las actividades definidas en el proceso de mejora (ver Sección 4.4) aplicadas al caso de estudio.

6.1. Descripción del problema

En esta sección se describe la realidad del caso de estudio al cual se aplica el proceso de mejora de calidad de datos. Este caso de estudio está inspirado en un trabajo realizado en el año 2012, por el Instituto de Computación (INCO) en el marco de un convenio con el Ministerio de Salud Pública[1](MSP), para evaluar la calidad de datos de algunos de los sistemas de información del ministerio. Por motivos de confidencialidad es necesario omitir información sensible, por lo que se presenta una pequeña porción anonimizada del caso de estudio original. A continuación, se hace una descripción de la realidad que existía al momento en que se llevó a cabo el convenio. Aunque se describe en tiempo presente, se debe tener en cuenta que corresponde a una situación de hace 10 años.

El certificado de nacido vivo (CNV) es un documento médico-legal que se expide a todo recién nacido con signos vitales y debe ser firmado por el médico o partera que atendió el parto. El CNV incluye los datos relativos a la identificación del certificado (número de certificado y fecha de la firma), nombres y apellidos del recién nacido, lugar e institución del nacimiento, datos de la madre (nombre, número del documento de identidad, país de nacimiento), datos del parto (día, hora y tipo) y los datos del recién nacido (sexo, peso, semana de gestación). Además, debe

[1]https://www.gub.uy/ministerio-salud-publica/home

incluir un número de cédula de identidad correlativo que administra la Dirección Nacional de Identificación Civil (DNIC)[2].

Existe un sistema informático (al que llamaremos SINV) que brinda el soporte al proceso de emisión del CNV electrónico (CNV-e). Este proceso se puede subdividir en dos etapas:

- Registro y actualización de los datos del recién nacido, de la madre, del parto y la firma por parte del médico o partida a cargo del nacimiento.
- La emisión del comprobante para su presentación en la DNIC.

El sistema utiliza un modelo de control de acceso basado en roles, en el cual cada rol tiene acceso a un subconjunto de funcionalidades. Por otro lado, el sistema SINV se comunica con servicios ofrecidos por la DNIC con el objetivo de validar los datos de la madre y reservar el número de cédula de identidad para el recién nacido.

Dada la importancia de la información que gestiona el sistema, la principal preocupación que existe actualmente se refiere a la calidad de los datos que se manejan. En particular, se entiende que existen problemáticas que pueden impactar negativamente en dicha calidad. Entre los principales problemas se destacan:

- **Problemáticas de los certificados en papel:** actualmente, la mayoría de los certificados se ingresan en el sistema informático. Existen excepciones donde debe realizarse en papel; por ejemplo, cuando el sistema SINV no está disponible o no se puede realizar la comunicación con servicios externos, como es el caso de la DNIC. Ante esta situación, los principales problemas radican en que los certificados en papel pueden contener datos no legibles. Esto lleva a que el digitador pueda ingresar al sistema un dato que no se corresponda con lo ingresado en el certificado en papel.
- **Problemáticas en los datos ingresados en los certificados:** Los principales problemas se dan ante la ausencia de datos que son solicitados y el ingreso de datos erróneos.

[2]https://dnic.minterior.gub.uy/

6.2. Aplicación del Proceso de mejora de la calidad de datos

El framework propuesto tiene tres componentes bien definidos: un proceso de mejora de calidad de datos, un modelo de madurez y un equipo de calidad de datos responsable de la aplicación del framework. Al tratarse de un caso de estudio, no puede evaluarse la inclusión en el framework del modelo de madurez y el equipo de calidad de datos.

La ejecución del proceso consiste en realizar las actividades propuestas en cada etapa y generar los documentos definidos. En la **Figura 4.4** se presentaron todas las etapas del proceso de calidad que será ejecutado. A continuación, se desarrollan los resultados de dicha ejecución.

Etapa: Definición del dominio.
El objetivo de esta etapa consiste en obtener una representación del dominio sobre el cual se aplica el proceso de mejora de calidad de datos. Esta etapa se compone de tres actividades:

1. Modelado de procesos de negocio y fuentes de datos.
2. Revisión de reglas, leyes y restricciones.
3. Relevamiento de tipos de usuario.

A continuación, se presenta el resultado de ejecutar las actividades mencionadas.

Actividad 1: Modelado de procesos de negocio y fuentes de datos.
Esta actividad consiste en describir los procesos de negocio de la organización y las entidades externas con las que se intercambian datos. Además, se deben identificar las fuentes de datos y su estructura detallando las tablas, campos y sus relaciones.

En el caso de estudio, una entidad externa implicada es la DNIC. La relación existente con el centro de salud consiste en el envío del número provisorio para la cédula de identidad del recién nacido. También los centros de salud realizan consultas a la DNIC para validar la información de la madre del recién nacido.

El principal proceso asociado al SINV es la emisión del CNV-e. En dicho proceso interviene personal administrativo del centro de salud, el médico/partero y la DNIC. En la **Figura 6.1** se presenta un modelo del proceso general de emisión de un CNV-e. La **Figura 6.2** describe de forma gráfica el sistema SINV.

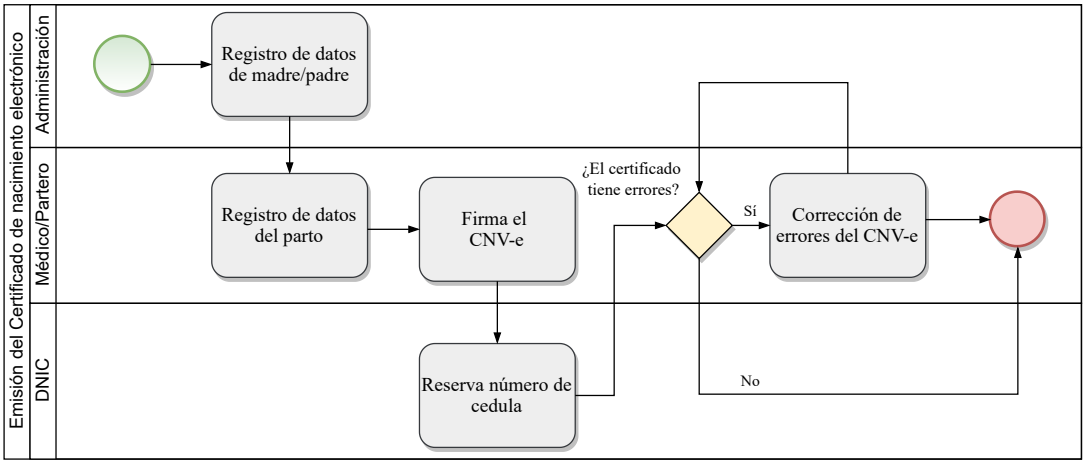


Figura 6.1: Proceso reducido de emisión de un CNV-e extraído de la documentación del convenio INCO-MSP.

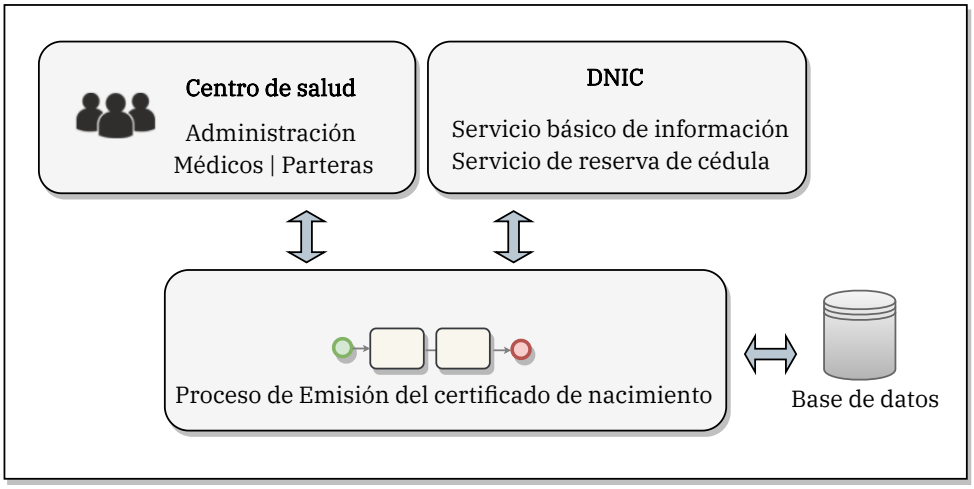


Figura 6.2: Resumen gráfico del sistema SINV extraído de la documentación del convenio INCO-MSP.

La fuente de datos del sistema SINV consiste en una base de datos que almacena la información de los CNV-e. En la **Figura 6.3** se presenta el modelo de datos simplificado del sistema SINV, que incluye cuatro tablas y sus correspondientes relaciones.

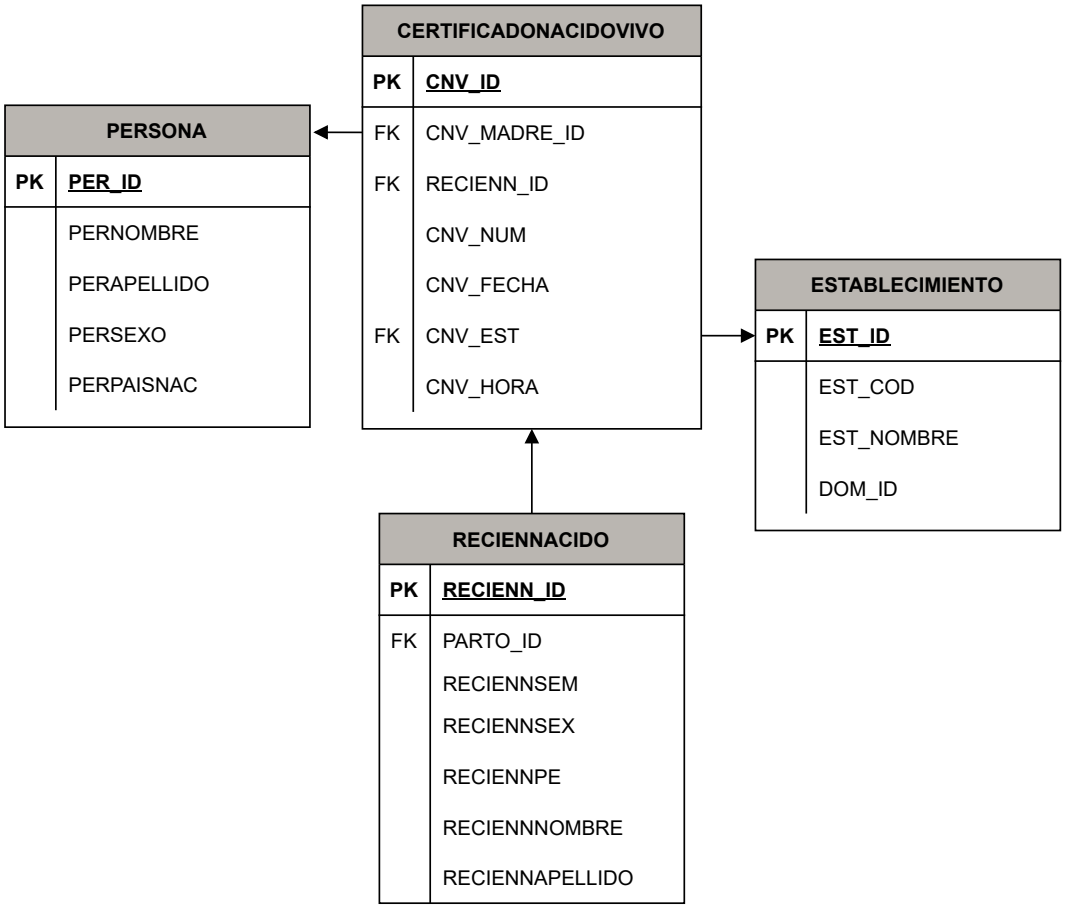


Figura 6.3: Modelo reducido de fuentes de datos extraído de la documentación del convenio INCO-MSP.

La tabla CERTIFICADONACIDOVIVO es la tabla central de los CNV-e. A partir de esta tabla se pueden encontrar los datos de los certificados ingresados en el sistema. La tabla PERSONA almacena los datos de quienes participan en este proceso (madre, partera, médico, etc). La tabla RECIENNACIDO contiene los datos vinculados al recién nacido. Por último, la tabla ESTABLECIMIENTO registra los centros de salud disponibles.

Actividad 2: Revisión de reglas, leyes y restricciones.
Esta actividad consiste en identificar restricciones internas o externas al SINV que puedan derivar en requisitos de calidad de datos. En la **Tabla 6.1** se presentan las reglas, leyes y restricciones identificadas.

Identificador	Descripción
#L1	Ley N. º 18331: Ley de protección de datos personales.
#L2	La cédula de identidad debe ser válida.

Tabla 6.1: Relevamiento de reglas, leyes y restricciones relevantes para el caso de estudio.

Actividad 3: Relevamiento de tipos de usuario.
Esta actividad consiste en identificar los usuarios internos de los centros de salud que manipulan datos. En la **Tabla 6.2** se presentan los usuarios identificados y los procesos con los que interactúan.

Identificador	Nombre	Descripción	Procesos con los que interactúa
#U1	Administrador	Es el usuario encargado de iniciar el trámite para la emisión del CNV.	Emisión del certificado de nacimiento electrónico.
#U2	Médico/Partera	Usuario encargado de registrar los datos del parto y firmar el CNV.	Emisión del certificado de nacimiento electrónico.

Tabla 6.2: Tabla de tipos de usuario del sistema SINV.

Etapa: Análisis
La etapa de análisis tiene como objetivo obtener una primera aproximación al conocimiento de los datos e investigar los problemas de calidad de datos presentes en el sistema SINV. Las actividades incluidas en esta etapa son:

1. Realizar caracterización de datos.
2. Identificación de problemas de calidad de datos.
3. Relevamiento de requisitos de calidad de datos.

A continuación, se presenta el resultado de ejecutar las actividades mencionadas.

Actividad 1: Realizar caracterización de datos

En esta actividad se deben analizar las características y propiedades más destacables de los datos objetivos. Es decir, se centra en el análisis de los atributos, su contenido y estructura, entre otros. En el caso de estudio se realizó únicamente una caracterización resumida de algunas variables relevantes debido a la limitante de no contar con acceso a los datos. La caracterización de una variable refiere a la información relativa a la variable que permite conocer su ubicación, tipo de dato y restricciones definidas. En las **Tablas 6.3**, **6.4** y **6.5** se presenta la caracterización de tres variables de interés extraída de la documentación del convenio INCO-MSP.

Propiedad	**Valor**
Tabla en BD	CERTIFICADONACIDOVIVO
Nombre de atributo en BD	CNVMADRE_ID
Tipo de atributo	VARCHAR(15)
Restricciones sobre atributo en BD	No nulo

Tabla 6.3: Caracterización de la variable cédula de la madre extraído de la documentación del convenio INCO-MSP.

Propiedad	**Valor**
Tabla en BD	RECIENNACIDO
Nombre de atributo en BD	RECIENNPE
Tipo de atributo	NUMBER(4,0)
Restricciones sobre atributo en BD	No nulo

Tabla 6.4: Caracterización de la variable peso del recién nacido extraído de la documentación del convenio INCO-MSP.

Propiedad	**Valor**
Tabla en BD	RECIENNACIDO
Nombre de atributo en BD	RECIENNSEX
Tipo de atributo	CHAR(1)
Restricciones sobre atributo en BD	No nulo

Tabla 6.5: Caracterización de la variable sexo del recién nacido extraído de la documentación del convenio INCO-MSP.

Actividad 2: Identificación de problemas de calidad de datos
En esta actividad se identifican y clasifican los problemas de calidad de datos de acuerdo a su gravedad. En esta etapa del proceso es habitual tener una primera aproximación a los problemas de calidad. Sin embargo, en este caso de estudio, al disponer de una caracterización de los datos muy superficial, no es posible identificar problemas de calidad de datos. Por lo tanto, en etapas posteriores del proceso se retoma la identificación de problemas de calidad de datos.

Actividad 3: Relevamiento de requisitos de calidad de datos
En esta actividad se identifican, analizan y priorizan los requisitos de calidad de datos existentes en el sistema SINV. En la **Tabla 6.6** se presentan los requisitos relevados en el caso de estudio y que están relacionados con las variables de interés analizadas anteriormente (ver **Tablas 6.3, 6.4 y 6.5**). La identificación de requisitos se realiza en base a los documentos generados en etapas anteriores. Por ejemplo, el requisito #R2 surge de la restricción #L2 (ver **Tabla 6.1**).

Identificador	**Descripción**	**Prioridad**
#R1	El sexo del recién nacido no puede ser nulo, y su valor debe estar entre: F,M,D,S	Alta
#R2	El campo CNVMADRE_ID debe ser una cédula de identidad válida.	Media
#R3	El peso del recién nacido no puede ser nulo y debe ser mayor a 0.	Baja

Tabla 6.6: Requisitos de calidad de datos de variables relevantes del sistema SINV.

Etapa: Definición del modelo de calidad de datos
La tercera etapa del proceso tiene como objetivo definir el modelo de calidad de datos acorde a los problemas y requisitos identificados. Esta etapa se compone de dos actividades:

1. Definición del modelo de calidad de datos.
2. Implementación de métricas de calidad de datos.

A continuación, se presenta el resultado de ejecutar las actividades mencionadas.

Actividad 1: Definición del modelo de calidad de datos
El modelo de calidad de datos define las dimensiones, factores, métricas y métodos de medición orientados a la resolución de los problemas de calidad de datos identificados hasta el momento.

A continuación, se presentan las dimensiones y factores de calidad de datos, considerados en el convenio INCO-MSP, relacionadas con las tres variables presentadas anteriormente:

- **Dimensión Exactitud:** refiere a la veracidad del dato. Se define como la cercanía entre el valor v que está siendo evaluado y el valor v' que se considera como la representación correcta del mundo real. El factor considerado para esta dimensión es la **correctitud sintáctica** que refiere a que el dato este escrito correctamente.
- **Dimensión Completitud:** refiere a la existencia de todos los valores necesarios para la tarea a realizar. El factor considerado para esta dimensión es la **densidad** que hace referencia a la cantidad de valores faltantes para determinado dato.
- **Dimensión Consistencia:** refiere a la satisfacción de reglas semánticas definidas sobre los datos. El factor considerado para esta dimensión es la **integridad de dominio** que refiere a la satisfacción de reglas de dominio que deben cumplir los datos para un determinado atributo en una tabla de la base de datos.

Algunas de las métricas consideradas son generales y pueden ser aplicadas a diferentes variables, como son el caso de las métricas **Nulos** (dimensión: Completitud - factor: Densidad) y **FormatoCI** (dimensión: Exactitud - factor: Correctitud sintáctica).

Nulos()

- **Parámetro:** nombre de variable
- **Descripción:** verifica si la celda tiene valor nulo.
- **Tipo de resultado:** (0,1)
- **Granularidad:** celda

FormatoCI()

- **Parámetro:** nombre de variable
- **Descripción:** verifica que el número tenga entre 7 y 8 cifras y que el último sea el dígito verificador correcto. (La métrica está definida para los datos actuales, por lo tanto no deberían existir cédulas superiores a 9.999.999.)

- **Tipo de resultado:** (0,1)
- **Granularidad:** celda

En las **Tablas 6.7**, **6.8** y **6.9** se presentan las dimensiones, factores, métricas y métodos de medición asociados a las variables cédula de identidad de la madre, peso y sexo del recién nacido respectivamente.

Dimensión	**Factor**	**Métrica**	**Procedimiento de medición**
Exactitud	Correctitud sintáctica	Formato CI	Se realiza la consulta SQL utilizando la función LENGHT y se aplica el algoritmo del dígito verificador.
Completitud	Densidad	Nulos	Comparar con NULL o con ' '.

Tabla 6.7: Modelo de calidad de datos para la variable cédula de identidad de la madre.

Dimensión	**Factor**	**Métrica**	**Procedimiento de medición**
Consistencia	Integridad de dominio	Verificar que el valor este entre 0 y 8000 (gramos).	Se realiza la consulta SQL aplicando la métrica.
Completitud	Densidad	Nulos	Comparar con NULL.

Tabla 6.8: Modelo de calidad de datos para la variable peso del recién nacido.

Dimensión	**Factor**	**Métrica**	**Procedimiento de medición**
Exactitud	Correctitud sintáctica	Verifica que el valor esté dentro de los valores posibles 'I' (Indefinido), 'M' (Masculino), 'D' (Desconocido),'F' (Femenino)	Si el valor es igual a 'I', 'M', 'F', 'M' o 'D', devolver 1, si no, devolver 0.
Completitud	Densidad	Nulos	Comparar con NULL o con ' '.

Tabla 6.9: Modelo de calidad de datos para la variable sexo del recién nacido.

Actividad 2: Implementación de métricas de calidad de datos

La implementación de las métricas definidas se realiza a través de consultas SQL. A continuación, se presenta la implementación de las métricas, extraídas de la documentación del convenio INCO-MSP, para las variables analizadas.

Variable cédula de identidad de la madre.

- **Correctitud sintáctica:** para verificar la validez de la CI se realizan dos acciones. En primer lugar, se realiza una consulta SQL en donde se contabiliza la cantidad de CI que cumplen con el largo establecido y luego se ejecuta un algoritmo que chequea si el documento es válido. La consulta SQL está dada por:

```
SELECT COUNT(*)
FROM CERTIFICADONACIDOVIVO
WHERE LENGTH(CNV_MADRE_ID) > 6 AND LENGTH(CNV_MADRE_ID) < 9
```

- **Densidad:** para verificar la densidad, se contabiliza la cantidad de registros con valores nulos. En el caso de estudio se menciona que, en muchos casos, los campos de tipo *varchar* tienen como valor un *string* vacío en lugar de NULL. Es por ello que se verifica cuantos registros existen tanto con valor NULL como *string* vacío (‘’), utilizando la siguiente consulta SQL:

```
SELECT COUNT(*)
FROM CERTIFICADONACIDOVIVO
WHERE CNV_MADRE_ID IS NULL OR CNV_MADRE_ID = ‘’
```

Variable peso del recien nacido.

- **Integridad del dominio:** verifica que los valores correspondientes al peso se encuentren dentro del rango establecido (el peso sea mayor a 0 pero no supere el valor 8000). La siguiente consulta permite saber cuántos registros pertenecen a este rango:

```
SELECT COUNT(*)
FROM RECIENNACIDO
WHERE RECIENNPE BETWEEN 1 AND 8000
```

- **Densidad:** para verificar la densidad, se contabiliza la cantidad de registros con valores nulos. La consulta SQL asociada es:

```
SELECT COUNT(*)
FROM RECIENNACIDO
WHERE RECIENNPE IS NULL
```

Variable sexo del recién nacido.

- **Correctitud sintáctica:** Se verifica que los valores correspondientes al sexo estén dentro de los valores posibles. La siguiente consulta nos permite saber los diferentes valores que adquiere la variable correspondiente al sexo del recién nacido y la cantidad de registros para cada uno.

```
SELECT RECIENNSEX, COUNT(*)
FROM RECIENNACIDO
GROUP BY RECIENNSEX
```

- **Densidad:** para verificar la densidad, se contabiliza la cantidad de registros con valores nulos. A continuación, se presenta la consulta SQL correspondiente:

```
SELECT COUNT(*)
FROM RECIENNACIDO
WHERE RECIENNSEX IS NULL OR RECIENNSEX = ‘’
```

Etapa: Evaluación de la calidad de datos

La cuarta etapa del proceso consiste en la ejecución de las medidas de calidad de datos. El objetivo es determinar el nivel de calidad de los datos del sistema SINV en función del modelo de calidad definido anteriormente. Las actividades que persiguen este objetivo son:

1. Medición de calidad de datos.
2. Evaluación de los problemas de calidad de datos.

A continuación, se presenta el resultado de ejecutar las actividades mencionadas.

Actividad 1: Medición de calidad de datos

En este caso de estudio se cuenta únicamente con la medición para el campo RECIENNSEX. En la **Tabla 6.10** se presentan la mediciones realizadas para esta variable. Los resultados de las mediciones fueron extraídos de la documentación del convenio INCO-MSP.

Métrica	**Resultado**	**Descripción**
Nulos(RECIENNSEX)	0	No existen registros con valor nulo en el campo correspondiente al sexo. Se concluye que la densidad para esta variable es 100 %
Verificación de datos	Ver **Tabla 6.11**.	De los 45683 registros en la tabla, el 99.9 % de ellos están correctos sintácticamente, mientras que existen 50 (0.1 %) que no están correctos, ya que el valor 'S' parece no ser válido.

Tabla 6.10: Mediciones de datos para la variable RECIENNSEX extraído de la documentación del convenio INCO-MSP.

Valor	**Cantidad registros**
M	23287
F	22376
I	11
D	9
S	50

Tabla 6.11: Resultados de los valores válidos para el campo sexo extraído de la documentación del convenio INCO-MSP.

Actividad 2: Re-evaluación de problemas de calidad de datos

En esta actividad se realiza una revisión de los problemas de calidad de datos existentes con el objetivo de identificar nuevos problemas de calidad que no fueron identificados en etapas anteriores. De acuerdo a la medición realizada en la actividad anterior, se identifica un único problema de calidad de datos. En la **Tabla 6.12** se presenta el problema mencionado.

Identificador	Descripción	Gravedad
#P1	El campo RECIENNSEX presenta valores inválidos.	Media

Tabla 6.12: Problemas de calidad de datos identificados.

Una vez concluidas las actividades de la etapa **Evaluación de la calidad de datos**, el proceso de mejora de calidad de datos brinda la posibilidad de redefinir el modelo de calidad, en caso de que se considere necesario. Esto consiste en volver a la etapa **Definición del modelo de calidad de datos**, realizar los cambios pertinentes en el modelo de calidad y obtener las medidas de calidad nuevamente. En caso de no querer modificar el modelo de calidad, se debe verificar si existen problemas de calidad de datos. En caso de no existir problemas en la calidad de los datos, se daría por finalizado el proceso. En cambio, si existen problemas de calidad se debe decidir si se quiere realizar mejoras. Si no se quisiera realizar mejoras de la calidad, se daría por finalizado el proceso. Por otro lado, si se quisiera realizar las mejoras de la calidad se continúa con la etapa **Mejora de la calidad de datos**. En este caso de estudio, luego de completar la etapa **Evaluación de la calidad de datos**, no se realizan cambios en el modelo de calidad, existen problemas de calidad de datos y se quieren solucionar, por lo tanto, se procede a continuar con la etapa **Mejora de la calidad de datos**. El flujo de la toma de decisiones se puede observar en la **Figura 6.4**.

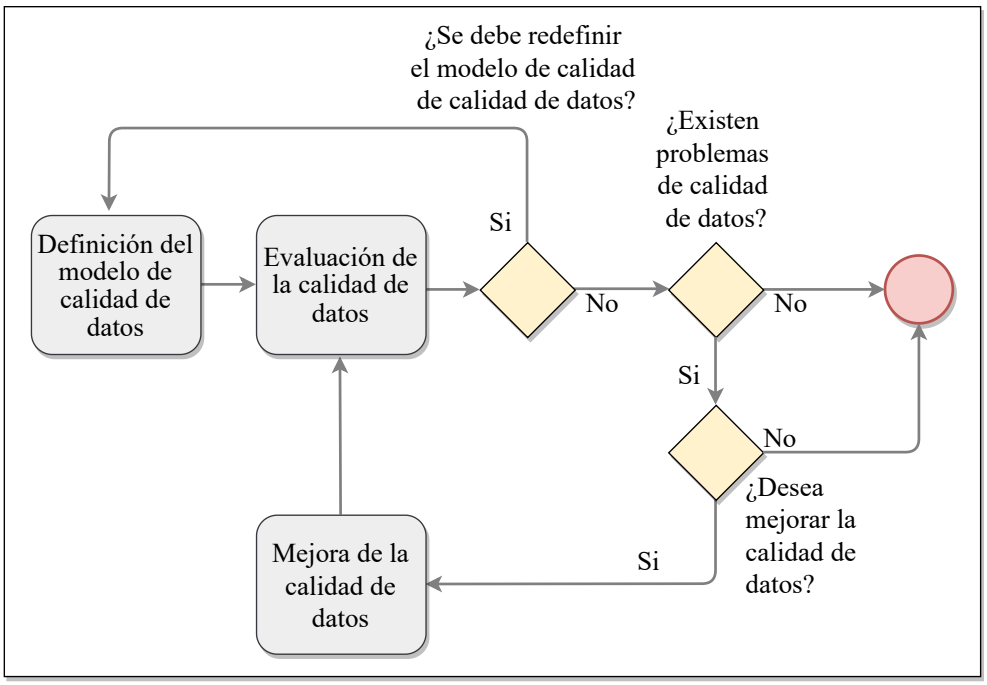


Figura 6.4: Flujo de toma de decisiones posterior a la etapa **Evaluación de la calidad de datos**.

Etapa: Mejora de la calidad de datos

La última etapa del proceso tiene como objetivo el diseño y gestión de estrategias enfocadas en mejorar la calidad de datos. Las actividades que definen esta etapa son:

1. Seleccionar estrategias y soluciones.
2. Ejecutar estrategias y soluciones.

A continuación, se presenta el resultado de ejecutar las actividades mencionadas.

Actividad 1: Seleccionar estrategias y soluciones

En esta actividad se define un plan de acción que depende de varios aspectos. Entre ellos se destacan los problemas de calidad de datos identificados, las tareas que son afectadas por dichos problemas, los usuarios de los sistemas que producen los datos, entre otros. En el caso de estudio, el plan de acción se define en función del problema identificado en la actividad anterior. En la **Tabla 6.13** se presenta este plan.

Identificador	**Estrategia**	**Descripción**	**Problema a resolver**	**Responsable**
#A1	Modificar sistema informático	Permitir únicamente el ingreso de valores válidos en el campo sexo del recién nacido	#P1	Área de Desarrollo

Tabla 6.13: Plan de acción para la variable sexo del recién nacido extraído de la documentación del convenio INCO-MSP.

Actividad 2: Ejecutar estrategias y soluciones

En esta actividad se ejecutan y documentan los resultados para cada estrategia definida en el plan de acción. La estrategia definida en la actividad anterior implica un cambio en el sistema. Ante esta situación no es posible verificar si dichas acciones contribuyeron en la mejora de la calidad de datos.

Una vez realizadas las actividades de mejora de calidad de datos, se vuelve a ejecutar la etapa **Evaluación de la calidad de datos** para confirmar si el plan de mejora tuvo el impacto deseado en la calidad de datos. En este caso de estudio, aunque sería de gran interés, no es posible ejecutar la etapa de mejora. Por lo tanto, se da por finalizada la ejecución del proceso.

Capítulo 7

Conclusiones y trabajo a futuro

El presente capítulo detalla las conclusiones obtenidas a partir del trabajo realizado y los lineamientos generales sobre posibles trabajos a futuro.

7.1. Conclusiones

El objetivo principal del proyecto fue proponer un framework para la gestión de calidad de datos de propósito general. Es decir, un framework que pudiera ser utilizado independientemente del dominio especifico de cada organización. Otro objetivo consistía en generar una documentación del framework definido, con el propósito de guiar su aplicación.

Antes de comenzar con la definición del framework, se estudiaron los principales fundamentos del área de calidad de datos, tomando como referencia el curso **"Calidad de datos e integración de datos"** brindado por el Centro de Posgrados y Actualización Profesional en Informática (CPAP)[1] a través de la Facultad de Ingeniería, UdelaR [2]. Por otro lado, se analizó bibliografía seleccionada sobre metodologías y frameworks existentes, que están enfocados en la gestión de calidad de datos. En dicho análisis se estudiaron los diferentes componentes de cada propuesta. Las metodologías y los frameworks analizados están orientados a dominios particulares, como por ejemplo salud o gobierno electrónico, entre otros. Sin embargo, todos incluyen una definición de calidad de datos, sus características más relevantes y actividades para la evaluación y mejora de la calidad de datos. Estas propuestas analizadas sirvieron como punto de partida para la creación del framework propuesto.

Como se mencionó antes, para la definición del framework se analizaron los componentes más relevantes de los frameworks y metodologías revisados. A partir de dicho análisis se decidió

[1] https://www.fing.edu.uy/es/cpap

la inclusión de tres componentes: un equipo responsable de la aplicación del framework, un modelo de madurez y un proceso de mejora de calidad de datos.

Contar con un equipo responsable garantiza la correcta aplicación del framework y aumenta las probabilidades de éxito del mismo [43]. En el framework propuesto se incluyeron los mismos roles del equipo propuesto por el framework de AGESIC [3], ya que se consideró que estos roles satisfacen las necesidades del framework.

Por otro lado, incluir un modelo de madurez facilita el conocimiento del estado actual de la organización permitiendo medir y conocer diferentes aspectos a mejorar. De las metodologías y frameworks analizados, únicamente el framework DQMP[4] incluye este tipo de modelo. Dadas las ventajas de contar con un modelo de madurez, se decidió incluir uno en el framework propuesto. Debido a que el modelo de madurez propuesto en el framework DQMP[4] está orientado al sector público de Estonia, se decidió incluir el modelo de madurez propuesto en el trabajo realizado por A. Kurniati y K. Surendro en [30]. Esta elección se basa principalmente en la simplicidad de su aplicación, su orientación al área de calidad de datos y su adaptabilidad a diferentes dominios.

Por último, se incluyó un proceso de mejora de calidad de datos. Este proceso de gestión se divide en etapas, donde cada etapa se compone de varias actividades. Uno de los grandes desafíos a la hora de definir este proceso fue el entendimiento de las tareas definidas en las metodologías y frameworks revisados. Al definir un proceso genérico, el principal desafío fue extraer las actividades genéricas excluyendo aquellas actividades particulares de los dominios. A grandes rasgos, el proceso definido se encarga de obtener toda la información sobre la organización, para luego realizar un análisis de los datos objetivo, identificando y priorizando los problemas y requisitos de calidad de datos. En función del análisis realizado, se define un modelo de calidad de datos acorde y se evalúa la calidad de los datos. Una vez seleccionados los problemas a resolver, se definen y aplican las estrategias de mejora de la calidad de datos adecuadas para cada problema. El valor agregado del proceso definido, respecto a lo analizado, consiste en la inclusión de las actividades más relevantes que permiten la mejora de la calidad de los datos independientemente del dominio en el cual se aplican.

Una vez implementado el framework, se generó su documentación utilizando la herramienta *EPF Composer*. En esta herramienta se documenta el proceso de mejora de calidad de datos, facilitando el acceso a la información de todos los elementos incluidos en el framework (roles, actividades, productos, etc). Contar con documentación del proceso de mejora facilita su apren-

dizaje y aplicación, así como también la institucionalización del mismo.

Por último, se utilizó un caso de estudio real para mostrar la aplicación del proceso de mejora propuesto. Al trabajar con un caso de estudio real, fue necesario reservar los procesos y estructura de la organización por motivos de confidencialidad, por lo que se debió acotar la realidad del caso. Otra limitante fue el no poder acceder a los datos, lo cual impidió definir y aplicar las métricas de medición, para luego seleccionar las estrategias y ejecutar el plan de acción. Este punto es muy importante, ya que si bien fue posible mostrar cómo ejecutar el proceso de calidad, no fue posible validar la efectividad real del framework completo. Sin embargo, durante la aplicación del proceso en el caso de estudio surgieron mejoras en algunas de sus etapas y actividades. Por lo tanto, consideramos que el caso de estudio presentado aportó valor al proyecto realizado.

Para finalizar, si bien se identificaron varios aspectos a tener en cuenta para trabajos futuros, se considera que se cumplieron los objetivos definidos al comienzo del proyecto. Dada la importancia de los datos en las organizaciones, se evidencia la necesidad de contar con herramientas que faciliten la gestión de la calidad de datos. Por esto, el framework propuesto apunta a que las organizaciones puedan adoptar una forma de trabajo que permita que los datos sean de mayor calidad. Esto implica mejorar y optimizar la toma de decisiones y poder mejorar los servicios brindados.

7.2. Trabajo a futuro

A continuación, se detallan algunas lineas de trabajo a futuro que podrían considerarse en pos de mejorar el framework propuesto.

- La calidad de datos es denominada *fitnesss for use* por la bibliografía de referencia de la comunidad de calidad de datos. Esto implica que la calidad de datos debe ser adecuada para su uso, es decir, valores de calidad que son aceptables para un uso, podrían no serlo para otro. Por lo tanto, esto implica incorporar al proceso el contexto de los datos, que es lo que aporta información sobre cómo, quiénes, cuándo y dónde son usados los datos. Por una cuestión de alcance no fue incluido en este proyecto, sin embargo, la consideración del contexto aportaría gran valor a la medición y evaluación de la calidad de los datos. De acuerdo con Anind K. Dey [48], el contexto es cualquier información que se pueda utilizar para caracterizar la situación de una entidad. Una entidad puede ser una persona, lugar u objeto que se considere relevante para la interacción entre un usuario y una aplicación, incluidos el usuario y las propias aplicaciones.

- Dado que el caso de estudio es acotado, sería positivo aplicar el framework en un caso real en el cual se puedan acceder a los datos y estructura de la organización en su completitud. Esto permitiría validar la efectividad del framework completo para evaluar la etapa de mejora de la calidad de datos.

Apéndice A

Dimensiones de calidad de datos

El presente anexo detalla las dimensiones mas relevantes de la calidad de datos y, para algunos casos, sus factores más conocidos. En particular, las primeras cinco dimensiones expuestas son las propuestas por AGESIC [49]. El resto de las dimensiones son las definidas por la norma ISO 25012 [19][44]. Cabe destacar que algunas de las dimensiones propuestas por AGESIC coinciden con las definidas por la norma ISO. En [50] se pueden visualizar ejemplos de métricas para las primeras cinco dimensiones.

Exactitud	
Definición	Proximidad entre un valor de datos *v* y un valor de datos *v'*, considerado como la representación correcta del fenómeno del mundo real que v intenta representar.
Factores asociados	Correctitud semántica - Correctitud sintáctica - Precisión Exactitud posicional absoluta - Exactitud posicional relativa - Fidelidad

Tabla A.1: Exactitud.

Factores relacionados con la dimensión Exactitud:

- **Correctitud semántica:** es la proximidad entre el valor *v* de un atributo y su verdadero valor *v'*.
- **Correctitud sintáctica:** es la proximidad entre el valor *v* de un atributo y los elementos del dominio de definición de dicho atributo.
- **Precisión:** captura el grado de detalle que posee un dato que lo hace útil para un determinado uso o que permite discriminarlo de otros datos que no son exactamente iguales.

Consistencia	
Definición	Captura la satisfacción de las reglas semánticas definidas sobre un conjunto de entidades de negocio o de sus atributos.
Factores asociados	Integridad Inter-entidad - Integridad Intra-entidad - Integridad de dominio

Tabla A.2: Consistencia.

Factores relacionados con la dimensión Consistencia:

- **Integridad inter-entidad:** captura la satisfacción de reglas entre atributos de diferentes entidades de negocio.
- **Integridad intra-entidad:** captura la satisfacción de reglas entre atributos de una misma entidad.
- **Integridad de dominio:** captura la satisfacción de reglas sobre los valores posibles que puede tomar un atributo.

Completitud	
Definición	Captura la medida en que los datos contienen toda la información de interés.
Factores asociados	Cobertura - Densidad

Tabla A.3: Completitud.

Factores relacionados con la dimensión Completitud:

- **Cobertura:** captura la proporción entre la cantidad de entidades existentes en una determinada colección de datos, y el total de entidades que deberían existir en dicha colección.
- **Densidad:** captura la proporción entre la cantidad de instancias de atributo con valores no nulos y el total de instancias de dicho atributo. Un valor nulo de una instancia de atributo A de una entidad E puede interpretarse de varias maneras:
  - E no posee A.

- Se desconoce si E posee A o no.
- E posee A pero se desconoce su valor.

Frescura	
Definición	Captura la rapidez con la que los cambios en el mundo real son reflejados en la actualización de los datos. La frescura es un tipo de exactitud no-estructural dependiente de la variable tiempo..
Factores asociados	Actualidad - Oportunidad

Tabla A.4: Frescura.

Factores relacionados con la dimensión Frescura:

- **Actualidad:** captura el tiempo de demora entre un cambio en el mundo real y la correspondiente actualización de los datos.
- **Oportunidad:** captura la demora que existe entre la actualización de un dato y el momento en el que éste se encuentra disponible para ser utilizado.

Unicidad	
Definición	Captura el grado en el que un dato del mundo real es representado de forma única.
Factores asociados	No duplicación - No contradicción

Tabla A.5: Unicidad.

Factores relacionados con la dimensión Unicidad:

- **No duplicación:** captura el grado de duplicación (o repetición) de un mismo dato.
- **No contradicción:** captura el grado de duplicación de una misma instancia de entidad del mundo real que es representada con datos que son contradictorios.

Comprensibilidad	
Definición	Grado en el que los datos tienen atributos que permiten ser leídos e interpretados por los usuarios y son expresados utilizando lenguajes, símbolos y unidades apropiados en un contexto de uso específico. Cierta información sobre la comprensibilidad puede ser expresada mediante metadatos.

Tabla A.6: Comprensibilidad.

Portabilidad	
Definición	Grado en el que los datos tienen atributos que les permiten ser instalados, reemplazados o eliminados de un sistema a otro, preservando el nivel de calidad en un contexto de uso específico.

Tabla A.7: Portabilidad.

Actualidad	
Definición	Grado en el que los datos tienen atributos que tienen la edad correcta en un contexto de uso específico

Tabla A.8: Actualidad.

Trazabilidad	
Definición	Grado en el que los datos tienen atributos que proporcionan un camino de acceso auditado a los datos o cualquier otro cambio realizado sobre los datos en un contexto de uso específico.

Tabla A.9: Trazabilidad.

Confidencialidad	
Definición	Grado en el que los datos tienen atributos que aseguran que los datos son sólo accedidos e interpretados por usuarios autorizados en un contexto de uso específico.

Tabla A.10: Confidencialidad.

Disponibilidad	
Definición	Grado en el que los datos tienen atributos que permiten ser obtenidos por usuarios y/o aplicaciones autorizados en un contexto de uso específico.

Tabla A.11: Disponibilidad.

Precisión	
Definición	Grado en el que los datos tienen atributos que son exactos o proporcionan discernimiento en un contexto de uso específico.

Tabla A.12: Precisión.

Recuperabilidad	
Definición	Grado en el que los datos tienen atributos que permiten mantener y preservar un nivel específico de operaciones y calidad, incluso en caso de fallos, en un contexto de uso específico.

Tabla A.13: Recuperabilidad.

Eficiencia	
Definición	Grado en el que los datos tienen atributos que pueden ser procesados y proporcionados con los niveles de rendimiento esperados mediante el uso de cantidades y tipos adecuados de recursos en un contexto de uso específico.

Tabla A.14: Eficiencia.

Accesibilidad	
Definición	Grado en el que los datos pueden ser accedidos en un contexto específico, particularmente por personas que necesiten tecnologías de apoyo o una configuración especial por algún tipo de discapacidad.

Tabla A.15: Accesibilidad.

Credibilidad	
Definición	Grado en el que los datos tienen atributos que se consideran ciertos y creíbles en un contexto de uso específico. La credibilidad incluye el concepto de autenticidad (la veracidad de los orígenes de datos, atribuciones, compromisos).

Tabla A.16: Credibilidad.

Conformidad	
Definición	Grado en el que los datos tienen atributos que se adhieren a estándares, convenciones o normativas vigentes y reglas similares referentes a la calidad de datos en un contexto de uso específico.

Tabla A.17: Conformidad.

Bibliografía

[1] I. S. P. Alvarez, "Minería de calidad de datos: Aplicación de técnicas de minería de datos para la evaluación de la calidad de los datos," 2018.

[2] Instituto de Computación. CPAP. Facultad de Ingeniería, Universidad de la República. (2020) Curso de calidad e integración de datos.

[3] AGESIC, "Framework para la Gestión de la calidad de Datos en Gobierno Digital," 2019, Accessed: 2022-06-12. [Online]. Available: https://bit.ly/2YCS4gB

[4] J. Tepandi, M. Lauk, J. Linros, P. Raspel, G. Piho, I. Pappel, and D. Draheim, "The Data Quality Framework for the Estonian Public Sector and Its Evaluation," ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2017.

[5] Central Statistical Agency (CSA), "Ethiopian data quality assessment framework," 2011.

[6] Canadian Institute for Health Information, "CIHI's Information Quality Framework," 2017.

[7] Health Information and Quality Authority, "Background paper to support guidance for a data quality framework for health and social care," 2018.

[8] F. Serra, "Context in data quality management," 2022, Tesis de doctorado en curso. Pedeciba Informática, Universidad de la República, Uruguay y Universidad de Tours, Francia.

[9] The Eclipse Foundation. (2018) Eclipse process framework project (EPF). Accessed: 2022-06-12. [Online]. Available: https://www.eclipse.org/epf/

[10] D. International, *DAMA-DMBOK: Data Management Body of Knowledge (2nd Edition)*. Denville, NJ, USA: Technics Publications, LLC, 2017.

[11] Instituto de Computación. Facultad de Ingeniería. Universidad de la República. (2021) Curso: Calidad de datos e información. [Online]. Available: https://eva.fing.edu.uy/course/view.php?id=1073

[12] C. Fox, A. Levitin, and T. Redman, "The notion of data and its quality dimensions," *Information Processing & Management*, 1994.

[13] H. Henderson, *Encyclopedia of Computer Science and Technology.* Facts on File, Inc, 2009.

[14] R. Y. Wang, "A product perspective on total data quality management," *Communications of the ACM*, vol. 41, no. 2, Feb. 1998.

[15] ISO, "ISO / IEC 25012: 2008, Ingeniería de software - Requisitos de calidad y evaluación de productos de software (SQuaRE) - modelo de calidad de datos," 2008, Accessed: 2021-05-03. [Online]. Available: https://bit.ly/3wKEZOY

[16] C. Batini and M. Scannapieco, *Data and Information Quality: Dimensions, Principles and Techniques*, 1st ed. Springer Publishing Company, Incorporated, 2016.

[17] B. Otto, Y. Lee, and I. Caballero, "Information and data quality in business networking: A key concept for enterprises in its early stages of development," *Electronic Markets*, vol. 21, Jun. 2011.

[18] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, Jul. 2009.

[19] J. Calabrese, S. Esponda, A. C. Pasini, M. Boracchia, and P. M. Pesado, "Guía para evaluar calidad de datos basada en ISO/IEC 25012," 2019.

[20] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, Mar. 1996.

[21] D. Strong, Y. Lee, and R. Wang, "10 potholes in the road to information quality," *Computer*, vol. 30, no. 8, 1997.

[22] L. Pipino, Y. Lee, and R. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, 07 2003.

[23] M Lebied, "Guide to data quality management & metrics for data control," 2018, Accessed: 2022-06-12. [Online]. Available: https://www.datapine.com/blog/data-quality-management-and-metrics/

[24] X. Mao, B. Gong, F. Su, K. Xu, K. Xian, D. Liu, and H. Guo, "Data Quality Management and Measurement." Springer Singapore, 2019.

[25] Instituto de Computación. Facultad de Ingeniería, Universidad de la República, "Calidad de datos e información - OpenFING," 2021, Accessed: 2022-06-12. [Online]. Available: https://open.fing.edu.uy/courses/cdi

[26] Talend. (2022) What is data cleansing? guide to data cleansing tools, services and strategy. Accessed: 2022-03-01. [Online]. Available: https://www.talend.com/resources/what-is-data-cleansing/

[27] P. Paygude and P. Devale, "Automated data validation testing tool for data migration quality assurance," *International Journal of Modern Engineering Research*, 2013.

[28] M. C. Paulk, C. V. Weber, S. M. Garcia, M. B. Chrissis, and M. Bush, "Key Practices of the Capability Maturity Model, Version 1.1:," Defense Technical Information Center, Fort Belvoir, VA, Tech. Rep., Feb. 1993.

[29] M. Ofner, B. Otto, and H. Oesterle, "A maturity model for enterprise data quality management," *Enterprise Modelling and Information Systems Architectures*, vol. 8, p. 2013, 12 2013.

[30] A. Kurniati and K. Surendro, "Designing IQMM as a Maturity Model for Information Quality Management," Jun. 2010.

[31] Acaps, "Data cleaning," 2016, Accessed: 2022-06-03. [Online]. Available: https://bit.ly/3iIMhME

[32] F. Naumann, "Data profiling revisited," *ACM SIGMOD Record*, 2014.

[33] Wikipedia. Query plan. Accessed: 2022-06-12. [Online]. Available: https://bit.ly/34Fidht

[34] RAE, "Diccionario de la lengua española," 2021, Accessed: 2022-06-12. [Online]. Available: https://bit.ly/30jyQx6

[35] M. Othman, M. N. Ahmad, A. Suliman, N. H. Arshad, and S. S. Maidin, "COBIT principles to govern flood management," *International Journal of Disaster Risk Reduction*, vol. 9, Sep. 2014.

[36] U. Van Heesch, P. Avgeriou, and R. Hilliard, "A documentation framework for architecture decisions," 2012.

[37] C. Cichy and S. Rass, "An Overview of Data Quality Frameworks," *IEEE Access*, 2019.

[38] C. Batini, D. Barone, F. Cabitza, and S. Grega, "A Data Quality Methodology for Heterogeneous Data," *International Journal of Database Management Systems*, vol. 3, Feb. 2011.

[39] C. Batini, F. Cabitza, C. Cappiello, and C. Francalanci, "A comprehensive data quality methodology for web and structured data," *International Journal of Innovative Computing and Applications*, 2008.

[40] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: a methodology for information quality assessment," *Information & Management*, vol. 40, no. 2, Dec. 2002.

[41] "Total Information Quality Management Handbook 3300.1," Tech. Rep., 2003. [Online]. Available: https://www.hud.gov/sites/documents/33001CIOH.PDF

[42] P. Falorsi, S. Pallara, A. Pavone, A. Alessandroni, E. Massella, and M. Scannapieco., "Improving the Quality of Toponymic Data in the Italian Public Administration," *In Proceedings of the ICDT 03 Workshop on Data Quality in Cooperative Information Systems (DQCIS '03)*, 2003.

[43] A. Caro, A. Fuentes, and M. Soto, "Desarrollando sistemas de información centrados en la calidad de datos," *Ingeniare. Revista chilena de ingeniería*, vol. 21, Apr. 2013.

[44] ISO. (2011) ISO 25012. Accessed: 2021-09-14. [Online]. Available: https://bit.ly/3qq8ILU

[45] B. Otto, K. Hüner, and H. Oesterle, "Identification of Business Oriented Data Quality Metrics," *14th International Conference on Information Quality (ICIQ 2009)*, pp. 122–134, Sep. 2009.

[46] P. Haumer, "Eclipse process framework composer," 2007, Accessed: 2022-03-01. [Online]. Available: https://www.eclipse.org/epf/general/EPFComposerOverviewPart1.pdf

[47] The Eclipse Foundation. (2010) Eclipse Process Framework (EPF) Composer. Accessed: 2022-05-15. [Online]. Available: https://www.eclipse.org/epf/general/EPF_Installation_Tutorial_User_Manual.pdf

[48] A. K. Dey, "Understanding and Using Context," *Personal and Ubiquitous Computing*, 2001.

[49] AGESIC, "Marco de trabajo para la gestión de la calidad de datos en gobierno digital," 2019, Accessed: 2022-01-02. [Online]. Available: https://bit.ly/34rStVx

[50] Instituto de Computación. Facultad de Ingeniería, Universidad de la República, “Calidad de datos,” 2021, Accessed: 2022-01-15. [Online]. Available: https://www.fing.edu.uy/inco/grupos/lins/demos/dq-model/