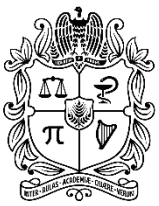


UNIVERSIDAD
NACIONAL
DE COLOMBIA

Modelo para la evaluación de la calidad de Datos Abiertos aplicado en el sector energético – caso Central Hidroeléctrica de Caldas

Mónica Rosa López Guayasamín

Universidad Nacional de Colombia
Facultad de Administración, Departamento de Informática y Computación
Manizales, Colombia
2021

Modelo para la evaluación de la calidad de Datos Abiertos aplicado en el sector energético – caso Central Hidroeléctrica de Caldas

Mónica Rosa López Guayasamín

Tesis presentada como requisito parcial para optar al título de:

Magister en Administración de Sistemas Informáticos

Director (a):

Ph.D. Néstor Darío Duque Méndez

Línea de Investigación:

Analítica de Datos

Grupo de Investigación:

Grupo de Ambientes Inteligentes Adaptativos GAIA

Universidad Nacional de Colombia

Facultad de Administración, Departamento de Informática y Computación

Ciudad, Colombia

2021

Dedicatoria

A mis Madre que es mi Ángel y de la cual siento su apoyo en todo momento. A mi esposo e hijo por la paciencia y el apoyo en esta difícil tarea…

A mi Asesor por confiar en mí y darme las fuerzas que necesitaba para cumplir con esta meta.

Agradecimientos

A la Universidad Nacional de Colombia Sede Manizales, Institución que me ha apoyado este proceso de investigación y a la Central Hidroeléctrica de Caldas CHEC S.A. E.S.P empresa que ha facilitado los datos, el conocimiento y los recursos para la realización de este trabajo.

Resumen

En Colombia la Ley 1712 del 2014 “Transparencia y Acceso a la Información”, define los datos abiertos como todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos.

En la política de Datos abiertos el tema central es la calidad de los mismos. Aunque existe varios enfoques y definiciones sobre el tema, no se encuentra un modelo único integrado con métricas específicas para poder evaluar objetivamente las características involucradas en la calidad. Mediante este proyecto de tesis de maestría se propone identificar las características relevantes en la calidad de los datos abiertos, definir métricas para su evaluación y mecanismos automáticos y/o semi-automáticos para su cálculo. La investigación que se llevará a cabo será teórica-práctica de tipo descriptivo ya que permitirá identificar las características que deben ser incorporadas en el modelo para la medición de la calidad de los datos abiertos.

A pesar de que el caso de aplicación es una compañía del sector energético en particular Central Hidroeléctrica de Caldas, la propuesta busca ser generalizada para diferentes ambientes de datos abiertos.

Palabras clave: Datos Abiertos, Calidad de datos, Métricas de Calidad

Model for the evaluation of the quality of Open Data applied to the energy sector - case of Central Hidroeléctrica de Caldas

Abstract

In Colombia, Law 1712 of 2014 "Transparency and Access to Information", defines open data as all primary or raw data, found in standard and interoperable formats that facilitate access and reuse, which are under the custody of public or private entities that fulfill public functions and that are made available to any citizen, freely and without restrictions, so that third parties can reuse them and create services derived from them.

In the Open Data policy, the central issue is the Quality of the same. Although there are several approaches and definitions on the subject, there is no single integrated model with specific metrics to be able to objectively evaluate the characteristics involved in quality. Through this master's thesis project, it is proposed to identify the relevant characteristics in the quality of open data, define metrics for their evaluation and automatic and / or semi-automatic mechanisms for their calculation. The research that will be carried out will be theoretical-practical of a descriptive type since it will allow identifying the characteristics that must be incorporated into the model for measuring the quality of open data.

Despite the fact that the application case is a company in the energy sector in particular Central Hidroeléctrica de Caldas, the proposal seeks to be generalized for different open data environments.

Keywords: Open Data, Data Quality, Quality Metrics

Contenido

Lista de Figuras

Lista de Tablas

1. Introducción

El Gobierno nacional aprobó en abril del 2018 el documento CONPES 3920 que define la política de explotación de datos (Big Data) para el Estado colombiano. Con este documento, el país asume el liderazgo regional al ser el primero en Latinoamérica, y octavo en el mundo, con una política pública integral que habilita el aprovechamiento de los datos para generar desarrollo social y económico (Republica, 2018) Esta política motiva la promoción del acceso público a la información, y el aumento significativo de datos oportunos y confiables.

La política de gobierno de Datos abiertos tiene definido los siguientes lineamientos:(Mintic, 2011)

- Servicios digitales de confianza
- Procesos internos seguros y eficientes a partir de las capacidades de TI
- Toma de decisiones basadas en datos
- Empoderamiento del ciudadano a través de la consolidación de un estado abierto
- Impulso en el desarrollo de territorios y ciudades inteligentes

Uno de los factores de alto impacto en la implementación de esta política, es la calidad de los datos, la cual representa un problema que está afectando los diferentes sectores del mercado: comerciales, oficiales, industriales, educativos entre otros. La falta de calidad de los datos en portales de datos abiertos puede manifestarse de varias formas:

Información inexacta, falta de integridad en los conjuntos de datos oficiales, datos obsoletos o conjuntos de datos corruptos (Sadiq & Indulska, 2017).

Para enfrentar esta situación, aunque existen varios enfoques que referencian el problema (Vetrò et al., 2016a), no se encuentra un modelo único integrado con métricas específicas para poder determinar cuáles son las características más importantes que deberán ser incorporadas en la implementación del diseño para la evaluación de calidad de datos publicados en los portales de datos abiertos. Para citar algunos casos, se mencionan los siguientes artículos que relacionan dicho tema:

- (Sadiq & Indulska, 2017) Menciona los desafíos de tratar calidad de datos con datos abiertos.
- (Kubler et al., 2018) Menciona que las organizaciones de hoy no prestan suficiente atención a la gestión del conjunto de datos.
- (Vetrò et al., 2016b) Al no publicar datos con calidad se pone en peligro la reutilización de los datos por parte del ciudadano.
- (Giovannini, n.d.) Se mencionan dimensiones de calidad que deberían ser utilizadas para los procesos de calidad de datos en portales de datos abiertos.
- (D´Agostino et al., 2018) Se plantea una estrategia de gobernanza de datos de salud con herramientas, conceptos y recomendaciones que permitirán a los países generar datos abiertos y de mayor calidad, así como confiables y seguros
- (Colborne & Smit, 2018) Se relacionan los riesgos que se presentan de calidad, integración y autenticidad de los datos abiertos.

Uno de los problemas relevantes en los sistemas de información de las empresas, es que, aunque permiten almacenar información asociada a un tema específico; no se ha dedicado mucho tiempo a implementar herramientas o validaciones que permitan garantizar la calidad de dicha información. Es por esto, que al momento de consolidar la información en repositorios soportados en TIC como Datamarts, Data-lake entre otros; se encuentra un problema mayor es que la información no pasa por procesos de calidad. Para complementar lo anterior, si la información que se utiliza para ser consolidada y

publicada en los portales de datos abiertos es tomada de repositorios sin pasar por herramientas de calidad, se tendrá información no confiable y posiblemente insuficiente para la toma de decisiones de los ciudadanos.

Mediante esta tesis de maestría se propone identificar las características relevantes que permitan evaluar de manera clara y consistente la calidad de los datos abiertos, definir métricas que se puedan incorporar en un modelo para su evaluación y mecanismos automáticos o semi-automáticos para su cálculo.

1.1 Objetivos

Objetivo General

Diseñar un modelo que incorpore características y componentes que permitan la evaluación de la calidad de los datos en los portales de Datos Abiertos aplicado al sector energético – caso Central Hidroeléctrica de Caldas que en adelante se denominará CHEC.

Objetivos Específicos

- Determinar las características a ser incorporadas en el diseño del modelo para la medición de la calidad de los datos abiertos en el sector energético
- Diseñar métricas automáticas y/o semi-automáticas para la medición de las características propuestas en el modelo.
- Construir el prototipo de modelo que permita evaluar la calidad de los datos en el portal de datos abiertos.
- Validar el modelo planteado sobre el portal de datos abiertos aplicado en el sector energético – caso CHEC.

1.2 Metodología

La investigación que se llevará a cabo será teórica-práctica de tipo descriptivo pues uno de los principales objetivos es identificar las características que deben ser incorporadas en el modelo para la medición de la calidad de los datos abiertos; de tal forma que se pueda identificar el comportamiento de estos datos. Por lo tanto, se debe realizar una revisión bibliográfica enmarcada en los conceptos de calidad asociada a los datos y cuyo alcance serán los portales de datos abiertos.

Una vez se realice el análisis anterior, se determinan las características relevantes a ser incluidas en la propuesta de Modelo. Posteriormente se procede a definir unas métricas para el cálculo automático y/o semi-automático de las características de calidad incluidas en el modelo de evaluación de datos de los portales de datos abiertos.

Mediante la revisión de una muestra de datos que permita recoger el comportamiento de varios portales de datos abiertos; se utiliza un enfoque cuantitativo permitiendo generalizar los resultados del estudio e identificar las posibles métricas que no han sido cubiertas por los modelos encontrados en el estudio.

1.3 Organización del documento

El desarrollo del documento comprende los siguientes capítulos:

El Capítulo 2 construye el marco teórico, y se presenta la definición de conceptos que serán mencionados en la presente tesis.

El Capítulo 3 define el estado del arte, donde se realiza una revisión de los trabajos previos de mayor importancia.

El Capítulo 4 consolida las variables de calidad para presentar la solución al problema planteado, donde se articula los temas investigados y definidos como atributos de calidad

a aplicar, el planteamiento de la alternativa de solución y el detalle de la implementación de la solución

El Capítulo 5 presenta la validación del Modelo, aplicando el prototipo con un caso práctico en la empresa Central Hidroeléctrica de Caldas y muestran los resultados de la validación.

Finalmente, el Capítulo 6 presenta conclusiones y recomendaciones resultado del trabajo que se realiza, y una propuesta de trabajos futuros que pueden servir como un punto de partida para continuar con esta temática.

2. Marco Teórico

2.1 Datos Abiertos

Los datos abiertos son información pública dispuesta en formatos que permiten su uso y reutilización bajo licencia abierta y sin restricciones legales para su aprovechamiento. En Colombia, la Ley 1712 de 2014 sobre Transparencia y Acceso a la Información Pública Nacional, define los datos abiertos en el numeral sexto como “todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos” (Rengifo et al., 2016)

2.2 Calidad de datos

Es la cualidad de un conjunto de información recogida en una base de datos, un sistema de información o un data warehouse que reúne entre sus atributos la exactitud, completitud, integridad, actualización, coherencia, relevancia, accesibilidad y confiabilidad necesarias para resultar útiles al procesamiento, análisis y cualquier otro fin que un usuario quiera darles (PowerData, n.d.).

Cuando se habla de calidad de datos se tiene que contemplar varios escenarios, el usuario final, el entorno empresarial, los procesos empresariales que de alguna u otra forma afectan estos datos.

En la Figura 2.1 se muestra un estándar universal de datos en dos capas y que puede ser aplicado a cualquier conjunto de datos, donde las dimensiones relevantes para dicho análisis son: Disponibilidad, Facilidad de uso, Confiabilidad, Importancia y Calidad de presentación (Zhu & Cai, 2015).

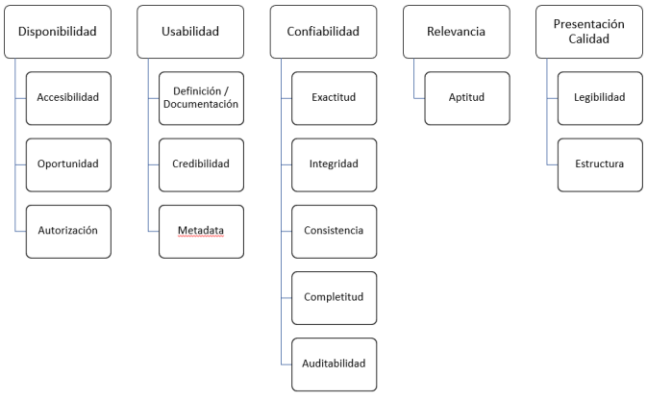


Figura 2.1 Un estándar universal de calidad de big data de dos capas para la evaluación. (Zhu & Cai, 2015)

- **Dimensión Disponibilidad**: Definido como el grado de conveniencia para que los usuarios puedan hacer uso de la información para lo cual es importante revisar criterios como accesibilidad, autorización y puntualidad.
- **Dimensión Facilidad de uso:** Se refiere a si los datos satisfacen las necesidades de los usuarios y son útiles para el público objetivo, esto incluye definición, documentación, confiabilidad.
- **Dimensión Confiabilidad:** Hace referencia a la confianza de estos datos, para lo cual atributos como precisión, consistencia, integridad, adecuación y audibilidad son relevantes a la hora de certificar dicho punto.
- **Dimensión Relevancia:** Es necesaria cuando se requiere validar la correlación entre el contenido de los datos y las necesidades de los usuarios.
- **Dimensión Calidad de presentación:** Permite a los usuarios entender los datos, para lo cual los atributos como legibilidad y estructura hacen parte de los criterios válidos para su medición.

Criterios calidad de datos. Los criterios se refieren al grado de cumplimiento de todos los requisitos necesarios para el propósito de los datos. Una de las metodologías aplicadas para definir estos criterios es (7+2) donde se mencionan 7 criterios medibles y dos documentables para el tema de calidad de datos (Morbey, 2013). En la Tabla 2.1 el autor (Morbey, 2013) presenta una forma de visualizar estos criterios de calidad y una breve descripción de estos.

Tabla 2.1 Criterios calidad de datos referenciado y traducido desde (Mobey, 2013)

	DQ Criterios de Calidad	Descripción
Medible automáticamente	(1) Completitud por fila (completitud horizontal)	¿Hay algún dato faltante o defectuoso en un registro? Todos los datos se ingresan de acuerdo con las necesidades comerciales.
	(2) Corrección sintáctica (conformidad)	¿Hay datos en un formato no estandarizado? Los datos encajan en el formato específico.
	(3) Ausencia de contradicciones (consistencia)	¿Qué valores de datos son contradictorios? Los datos no contradicen las especificaciones de integridad (reglas comerciales, valores empíricos) o rangos de valores definidos (dentro del grupo de datos, en comparación con otros grupos de datos, en el tiempo transcurrido).
	(4) Precisión incl. Actualidad	¿Qué datos son incorrectos o están caducados? Anotación correcta y actualizada (puntualidad) de nombres, direcciones, productos, etc.
	(5) Ausencia de repeticiones (libre de duplicados)	¿Qué registros de datos o contenidos de columnas se repiten? Sin duplicados (búsqueda de sinónimos y similitudes), sin homónimos, sin superposición (continuidad), todo es identificable con precisión (unicidad).
	(6) Integridad referencial empresarial (integridad)	¿Qué datos o relaciones de referencia faltan? No habrá ningún cliente sin contrato, se listarán los productos
	(7) Consistencia (sumas de verificación cruzada, integridad vertical)	¿Existe coherencia de datos en todos los sistemas? Por ejemplo: en una fecha señalada, el número de contratos en la fuente de datos es exactamente el mismo que el número de contratos en el DWH.
Documental	(8) Disponibilidad de documentación (capacidad de búsqueda)	¿Se pueden encontrar los datos de forma fácil y rápida (por ejemplo, utilizando funciones comunes de "búsqueda") ¿Están etiquetados los datos?
	(9) Consistencia normativa	Debe asegurarse que el nombre y el significado de ciertos datos sea el mismo en todos los sistemas, procesos y departamentos de la organización.

En la revisión del Estado del Arte se amplían los conceptos de acuerdo a las dimensiones y propuestas específicas por diferentes autores.

2.3 Leyes

Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional:

Estipulada en la Ley 1712 de 2014. plantea:

Artículo 3°. Otros principios de la transparencia y acceso a la información pública. En la interpretación del derecho de acceso a la información se deberá adoptar un criterio de razonabilidad y proporcionalidad, así como aplicar los siguientes principios:

Principio de transparencia, Principio de buena fe, Principio de facilitación, Principio de no discriminación, Principio de gratuidad, Principio de celeridad, Principio de eficacia, Principio de la calidad de la información(Congreso de la República, 2014).

2.4 Almacenamiento

El concepto de almacenamiento se considera importante en el trabajo de tesis ya que, para la gestión de calidad de datos, se requiere la consolidación y evaluación de la información de los datos en el tiempo, facilitando las actividades de monitoreo y control de los datos a los usuarios finales. Para dicha gestión en el tiempo y teniendo en cuenta el volumen de información es considerado que el tema de almacenamiento de la información en repositorios de datos como Datamart de calidad es una opción válida para complementar la solución que se propone en esta tesis.

Data Warehouse: Es un repositorio de datos que ofrece características especiales a sus datos como son: estable, coherente, confiable, histórica. Este concepto incluye detalles como:

- Orientado a un tema: Es una colección de información alrededor de un tema central.
- Integrado: Incluye múltiples orígenes de datos e información consistente
- Variable en el tiempo: Se tienen fotos de información donde la periodicidad es parametrizable por el analista y depende de las necesidades del usuario
- No volátil: Solo de lectura para usuarios finales.

Datamart: Un subconjunto de datos de la Data Warehouse cuyo objetivo es responder a un determinando análisis, función o necesidad dependiendo del usuario. Los datos se pueden estructurar en diferentes modelos como son: copo de nieve, estrella. Los conceptos asociados a este tema son:

- Hechos: Corresponde a los históricos de la información
- Dimensiones: Corresponde a la vista, como queremos visualizar esta información.

  - Modelo estrella: Consiste en una tabla de hechos y una o más dimensiones con las que se puede modelar la información. En la estructura de hechos se tienen medidas asociadas a esta información.
- Modelo copo de nieve: Es un modelo derivado del modelo estrella, donde la estructura dimensión se puede normalizar en múltiples tablas.

Data Lake: Un Data Lake es un repositorio en el que se almacenan todos en bruto, sin ninguna organización, para analizarlos posteriormente los datos de la empresa. Independientemente de que estén estructurados o no, todos estos se encuentran. De hecho, las empresas vierten los datos y los recuperan cuando quieren. Únicamente en ese momento se procede a ordenarlos y a diseñar una estructura que haga más fácil su posterior análisis (PowerDAta, n.d.).

Durante el desarrollo de un almacén de datos se gasta una cantidad considerable de tiempo analizando las fuentes, entendiendo los procesos de negocio y perfilando los datos. Como resultado, se obtiene un modelo de datos altamente estructurado que ya está listo para la generación de informes.

Una gran parte de este proceso incluye también un procedimiento de toma de decisiones. Qué datos se incluyen y cuáles no. Por lo general, si los datos no se utilizan para responder a preguntas específicas o no son imprescindibles en algunos informes pueden excluirse. De esta manera, se simplifica el modelo y se conserva el espacio.

Sin embargo, un Data Lake conserva todos los datos. No solo los que son vitales en ese preciso momento sino todos aquellos que están guardados que pueden hacer falta en algún momento. Lo que proporciona varios beneficios sobre todo en cuanto a la parte de análisis.

Una alternativa para entender el concepto se puede visualizar en el siguiente gráfico:

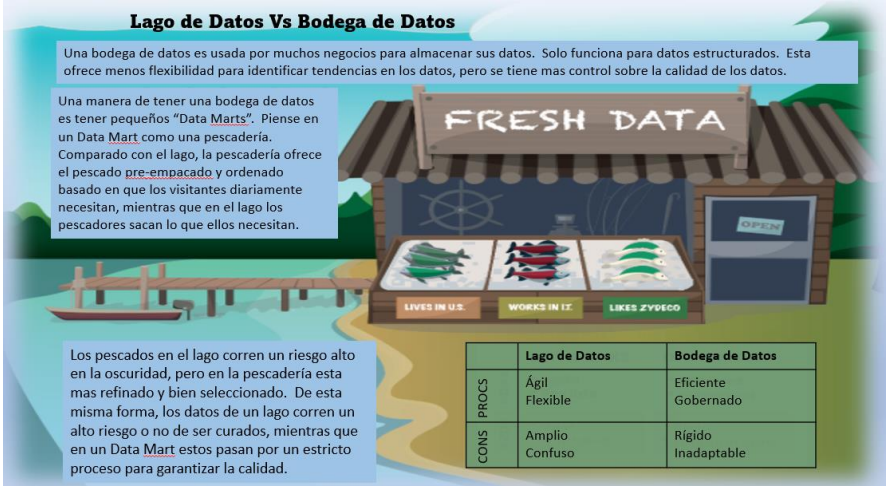


Figura 2.2 Diferencia Lago de Datos y Bodega de Datos (PowerData, n. d.)

ETL: Los procesos ETL son una parte de la integración de datos, pero es un elemento importante cuya función completa el resultado de todo el desarrollo de la cohesión de aplicaciones y sistemas. Todo proceso ETL consta de estas tres fases: extracción, transformación y carga (PowerData, 2017).

Para la fase de extracción de datos, se destacan los siguientes puntos:

- Extraer los datos desde los sistemas de origen.
- Analizar los datos extraídos obteniendo un chequeo.
- Interpretar este chequeo para verificar que los datos extraídos cumplen la pauta o estructura que se esperaba. Si no fuese así, los datos deberían ser rechazados.
- Convertir los datos a un formato preparado para iniciar el proceso de transformación

Es importante exigir siempre que esta tarea cause un impacto mínimo en el sistema de origen. Este requisito se basa en la práctica ya que, si los datos a extraer son muchos, el sistema de origen se podría ralentizar e incluso colapsar, provocando que no pudiera volver a ser utilizado con normalidad para su uso cotidiano. (PowerData, 2017).

En la fase de transformación se definen una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Estas directrices pueden ser declarativas, pueden basarse en excepciones o restricciones, pero para potenciar su pragmatismo y eficacia, hay que asegurarse de que sean: Declarativas, Independientes, Claras, Inteligibles, con una finalidad útil para el negocio

En la fase de carga, los datos son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes, sin embargo, se pueden identificar las siguientes acciones: (PowerData, 2017).

- **Acumulación simple:** Consiste en realizar un resumen de todas las transacciones comprendidas en el período de tiempo seleccionado y transportar el resultado como una única transacción hacia el data warehouse, almacenando un valor calculado que consistirá típicamente en un sumatorio o un promedio de la magnitud considerada.
- **Rolling:** Este proceso sería el más recomendable en los casos en que se busque mantener varios niveles de granularidad. Para ello se almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones de la unidad de tiempo o diferentes niveles jerárquicos en alguna o varias de las dimensiones de la magnitud almacenada (por ejemplo, totales diarios, totales semanales, totales mensuales, etc.).

3. Estado del Arte

El propósito de este aparte es presentar la importancia de un modelo que permita evaluar la calidad de los datos en los portales de datos abiertos, mediante los trabajos relacionados. Además, mostrar el espacio abierto donde se enmarca esta Tesis.

3.1 Revisión Sistemática

Los datos abiertos son una herramienta que le permite a las empresas y al ciudadano en general hacer uso de dichos recursos para la toma de decisiones. Para realizar este trabajo de investigación se aplicaron en diferentes etapas del trabajo de Tesis varias ecuaciones bibliográficas que permitieron identificar artículos que fueron importantes para el caso de estudio, los cuales se describen a continuación:

Selección Inicial: Se identifican 3 palabras relevantes las cuales son utilizadas con comodines especiales, y conectores los cuales se presentan en la Tabla 3.1.

Tabla 3.1 Ecuación bibliográfica inicial

COMPONENTES (AND)			
PALABRAS CLAVE (OR)	***Datos Abiertos***	***Calidad de datos***	***Metricas de Calidad de Datos***
	OpenData*	Data Quality	data*quality*metrics*
	open*data*	calidad*de*datos	Indicador*
	datos*abiertos*	data*quality*problems	metricas*calidad*datos
ECUACIÓN DE BÚSQUEDA			
(OpenData OR open*data* OR datos*abiertos*) AND (Data*Quality OR calidad*datos OR data*quality*problems) AND (data*quality*metrics OR indicador*calidad*datos OR metricas*calidad*datos)*			

Fuente: Elaboración propia
Para aumentar la relevancia de los resultados se aplica ademas de la cadena de búsqueda los siguientes criterios
Año de publicación: Entre 2015 y 2021
Tipo de publicación: Articulo Científico o libros
Idioma: Ingles, español o portugués

Aunque de esta ecuación se encuentran pocos artículos por la precisión y detalles de los elementos de búsqueda, los artículos que aquí se identifican son incluidos en el proceso

de investigación. Como consecuencia de este ejercicio se realiza otra búsqueda sistemática

Ecuación siguiente: Se conservan las 3 palabras relevantes las cuales son utilizadas con comodines especiales, solo que se minimizan las palabras y comodines a utilizar, esta ecuación se presenta en la Tabla 3.2.

Tabla 3.2 Ecuación bibliográfica alterna

COMPONENTES (AND)			
PALABRAS CLAVE (OR)	***Datos Abiertos***	***Calidad de datos***	***Metricas de Calidad de Datos***
	open*data	quality	metric*
	datos*abiertos	calidad	indicador*
		problems	
ECUACIÓN DE BÚSQUEDA			
*((open*data OR datos*abiertos) AND (quality OR calidad OR problems) OR (metric* OR indicador*))*			

Fuente: Elaboración propia
Para aumentar la relevancia de los resultados se aplica ademas de la cadena de búsqueda los siguientes criterios
Año de publicación: Mayor a 2015
Tipo de publicación: Articulo Científico o libros
Idioma: Ingles, español o portugués

En esta selección aparecen más documentos sin embargo se requiere una validación detallada adicional para poder identificar los artículos a ser incluidos en el proceso de investigación.

Ecuación final: Se conservan las 3 palabras relevantes las cuales son utilizadas con comodines especiales, se minimizan las palabras y comodines a utilizar, esta ecuación se presenta en la Tabla 3.3.

Tabla 3.3 Ecuación bibliográfica final

COMPONENTES (AND)			
PALABRAS CLAVE (OR)	***Datos Abiertos***	***Calidad de datos***	***Metricas de Calidad de Datos***
	OpenData	Quality	metric*
	datos abiertos	calidad	ind*
		problems	
ECUACIÓN DE BÚSQUEDA			
(((open data OR datos abiertos) AND (quality OR calidad OR problems) AND (metric OR ind*)))*			

Fuente: Elaboración propia
Para aumentar la relevancia de los resultados se aplica ademas de la cadena de búsqueda los siguientes criterios
Año de publicación: Mayor a 2015
Tipo de publicación: Articulo Científico o libros
Idioma: Ingles, español o portugués

Mediante esta ecuación se permiten identificar más artículos y aunque se requiere una mayor depuración en los artículos, con este ejercicio se pudieron encontrar más elementos de estudio que favorecieron el proceso investigativo.

Es importante mencionar que algunos de los recursos o bibliotecas digitales que fueron consultados son: IEEE, Springer, Web of Science, Science Direct, MINTIC, Scielo, Scopusc, entre otros.

De esta revisión bibliográfica sistemática, cuyo enfoque contenía los elementos de Datos Abiertos, Métricas y Calidad de datos, se visualiza en la Figura 3.1 un resumen de los tipos de temas abordados y la cantidad de artículos que fueron identificados por temáticas.

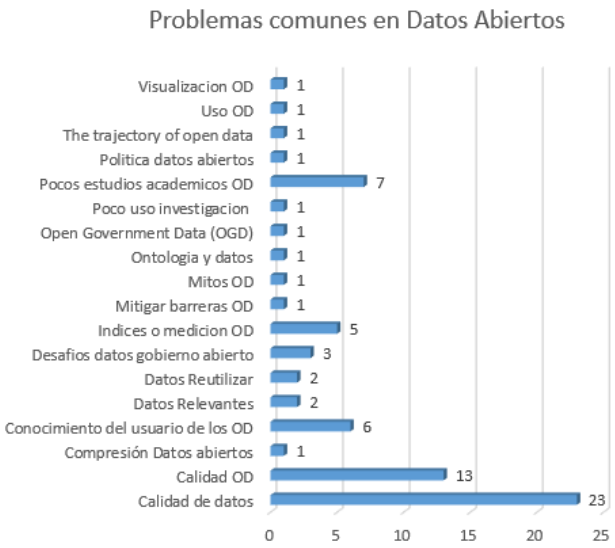


Figura 3.1 Problemas comunes en Datos Abiertos. Fuente: Elaboración Propia

Como se puede evidenciar en la Figura 3.1, la temática que se aborda con más frecuencia por los autores es Calidad de Datos publicados en los portales de Datos Abiertos. Teniendo en cuenta lo anterior y basados en la importancia de la Calidad de datos para las organizaciones este es el espacio de investigación que se trabajó en esta Tesis.

Basados en los resultados de la revisión bibliográfica sistemática, y una vez realizada una pre-clasificación de los tipos de las temáticas abordadas por los artículos se observa que hay algunas tendencias más abordadas por los autores de las cuales podemos citar algunas: Estudios académicos de Datos Abiertos (Verhulst & Young, 2017), (Santos & Guanaes, 2018), (Ríos Ramírez & Garro, 2018), (Bonina & Scrollini -Ilda, n.d.), (Scavuzzo et al., 2018), (Oliveira et al., 2017), Indicadores de Datos abiertos (Castillo, 2018), (Pane et al., n.d.), (Torchiano et al., 2017), Conocimiento del usuario de los datos abiertos (Machova & Lnenicka, 2017), (Concha & Naser, n.d.), (Danneels et al., 2017), (Ruijer et al., 2017), Mitigar barreras de Datos Abiertos (Beno et al., 2017), Trayectoria de Datos Abiertos (Sieber & Johnson, 2015).

Por otro lado, aunque la estrategia de publicar en los portales de las empresas los Datos Abiertos ha facilitado a la ciudadanía tener acceso a la información de interés en diferentes áreas como salud, educación, gobierno, entidades entre otros; son varios factores que toman importancia a la hora de publicar la información como son: Calidad de la información, la importancia de los datos para el ciudadano, el conocimiento de los usuarios de la información publicada, entre otros aspectos.

Si bien es cierto la publicación de los datos depende del interés de la entidad en dar a conocer información para los ciudadanos, hay un tema de cultura en las organizaciones que es necesario fomentar a nivel de altos directivos; no solo es un cumplimiento a los entes gubernamentales, sino entender la importancia de los datos abiertos para la comunidad y que puede llegar a ser una herramienta estratégica para generar valor basados en la información.

Es importante mencionar que la publicación de los datos abiertos en los portales, dan la posibilidad a los usuarios de generar nuevos productos, servicios y conocimiento en las diferentes áreas como son salud, educación, gobierno, entidades entre otros; por lo cual se considera relevante plantear alternativas para las empresas que permitan evaluar la calidad de estos datos antes de realizar la publicación de la información en dichos portales.

Como el espacio de investigación realizado por la Tesis es Calidad de Datos en los portales de Datos Abiertos se requiere determinar las dimensiones involucradas en los datos abiertos ya que existen diferentes autores que referencian algunas dimensiones, o las dimensiones tienen similitudes en su definición pero los autores los relacionan con nombres diferentes, por lo cual uno de los ejercicios siguientes es identificar los artículos que hablan del tema, e identificar las dimensiones que han sido más abordadas en los estudios, y aquellas que se consideran como problemas más frecuentes en los portales de datos abiertos. Basados en la revisión sistemática, en la figura 3.2 se presenta una clasificación de las dimensiones de calidad que se referencian como problemas más

frecuentes que se están presentando en los portales de datos abiertos, donde los valores referenciados corresponden a la cantidad de artículos que abordan estas dimensiones.

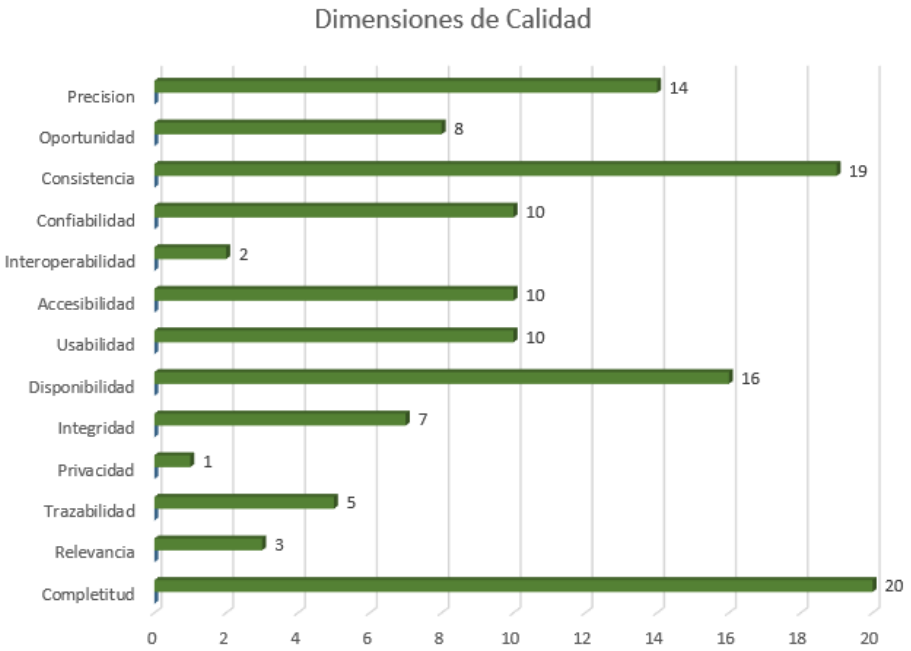


Figura 3.2 Dimensiones de Calidad en Datos Abiertos. Fuente: Elaboración Propia

Para complementar este ejercicio de identificación de artículos que están asociados a la temática de Calidad de datos como un componente importante en los datos abiertos, se detalla en la Tabla 3.4 un consolidado de las referencias bibliográficas que han abordado este concepto en las diferentes dimensiones asociadas a Calidad.

Tabla 3.4 Lista de Dimensiones Calidad de datos referenciadas por autores

DIMENSIONES CALIDAD DE DATOS	REFERENCIAS	CANTIDAD DE REFERENCIAS
Disponibilidad	(Sadiq & Indulska, 2017) - (Kubler et al., 2018) -(D´Agostino et al., 2018)- (Xia et al., 2018a) - (P. Zhang et al., 2018)-(Dawes et al., 2016) - (Safarov et al., 2017) - (Vetrò et al., 2016c) -	16

	(Abella & De-pablos-heredero, n.d.-a) - (Yoon et al., 2019) - (Bicevskis et al., 2018) - (Nikiforova, 2018) - (Utamachant & Anutariya, 2018) - (Reiche et al., 2014) - (Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC], 2019c) - (25000, 2021)	
Completitud	(Sadiq & Indulska, 2017) -(Kubler et al., 2018) - (Torchiano et al., 2017)- (Ahmed, 2018) -(Xia et al., 2018b) - (Utamachant & Anutariya, 2018) - (Kao et al., 2017) - (Ruijer et al., 2020) - (Vetrò et al., 2016c) - (R. Zhang et al., 2019)(Abella & De-pablos-heredero, n.d.-a) - (Daraio et al., 2016) - (Bicevskis et al., 2018) - (Nikiforova, 2018) - (Morbey, 2013) - (Utamachant & Anutariya, 2018) - (Behkamal et al., 2014) - (Reiche et al., 2014) - (Batini & Scannapieca, 2006) -(Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC], 2019b) - (Icontec, n.d.) - (ISO, 2021)	20
Consistencia	(Sadiq & Indulska, 2017)- (Talukder et al., 2018) - (Giovannini, n.d.) - (Torchiano et al., 2017) - (Ferney et al., 2018) - (Ahmed, 2018) - (Xia et al., 2018b) -(Utamachant & Anutariya, 2018) - (Vetrò et al., 2016c) -(Abella & De-pablos-heredero, n.d.-a) - (R. Zhang et al., 2019) - (Daraio et al., 2016) - (Bicevskis et al., 2018) - (Nikiforova, 2018) - (Morbey, 2013) - (Utamachant & Anutariya, 2018) - (Behkamal et al., 2014) - (Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC], 2019a) -(ISO, 2021)	19

Usabilidad	(Kubler et al., 2018)- (Giovannini, n.d.) - (P. Zhang et al., n.d.) -(Dawes et al., 2016) - (Safarov et al., 2017) - (Ruijer et al., 2020) - (Vetrò et al., 2016c) - (Abella & De-pablos-heredero, n.d.-b) - (ISO, 2021) -(Abella et al., 2019)	10
Oportunidad	(Giovannini, n.d.) - (Ferney et al., 2018) - (Giovannini, n.d.) (Safarov et al., 2017) - (Vetrò et al., 2016c) - (R. Zhang et al., 2019) - (Daraio et al., 2016) - (Bicevskis et al., 2018) - (Nikiforova, 2018)	8
Integridad	(D´Agostino et al., 2018) - (Ferney et al., 2018) - (P. Zhang et al., n.d.) - (Yoon et al., 2019) - (Bicevskis et al., 2018) - (Nikiforova, 2018) - (Behkamal et al., 2014)	7
Accesibilidad	(Kubler et al., 2018)- (Giovannini, n.d.) - (Ferney et al., 2018) - (P. Zhang et al., n.d.) - (Utamachant & Anutariya, 2018) - (Dawes et al., 2016) - (R. Zhang et al., 2019) - (Morbey, 2013) -(Reiche et al., 2014)	11
Confiabilidad	(Kubler et al., 2018) - (Talukder et al., 2018) - (Giovannini, n.d.) - (P. Zhang et al., n.d.) - (Utamachant & Anutariya, 2018) - (Yoon et al., 2019) - (Batini & Scannapieca, 2006) - (Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC], 2019a) -(ISO, 2021)	10
Relevancia	(Kubler et al., 2018)- (Ahmed, 2018) - (P. Zhang et al., n.d.)	3
Trazabilidad	(Ferney et al., 2018) - (Vetrò et al., 2016c) - (Batini & Scannapieca, 2006) - (Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC], 2019a) -(ISO, 2021)	5

Interoperabilidad	(Giovannini, n.d.) -(Dawes et al., 2016)	2
Precisión	(Brezočnik et al., 2018) - (Oliveira et al., 2017) - (Giovannini, n.d.) - (Torchiano et al., 2017) - (Ahmed, 2018) - (Xia et al., 2018b) - (Utamachant & Anutariya, 2018) -(Morbey, 2013) - (Behkamal et al., 2014)- (Reiche et al., 2014) - (Batini & Scannapieca, 2006) - (Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC], 2019a) -(ISO, 2021)	14
Actualidad	(Utamachant & Anutariya, 2018) - (Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC], 2019a) - (ISO, 2021)	3
Conformidad	(Ferney et al., 2018)- (Utamachant & Anutariya, 2018) - (Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC], 2019a) - (ISO, 2021)	4
Comprensibilidad	(Veljković et al., 2014) - (Utamachant & Anutariya, 2018) - (Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC], 2019a) - (ISO, 2021)	4

3.2 Encuesta

Uno de los puntos importantes a destacar como contribución a la Tesis es poder asociar la experiencia de empresa con el enfoque de la investigación a trabajar que en el caso particular es Calidad de datos abiertos. Para poder hacer esta asociación lo que se hizo fue diseñar un mecanismo que para el caso CHEC se define diseñar una encuesta, donde los usuarios que han trabajado con información en CHEC pudieran emitir sus

preferencias o clasificar de acuerdo a su vivencia cuales serían los puntos a considerar más importantes a tener en cuenta cuando se trata de Calidad.

Es importante destacar, que, para trabajar de manera coherente a la revisión sistemática, se parte de la identificación de dimensiones más relevantes o referenciadas por los autores, y teniendo en cuenta que el enfoque de este trabajo de Tesis era el diseño de un Modelo para la evaluación de datos abiertos en una empresa del sector; se considera importante identificar los siguientes aspectos:

- Cargo en la compañía. Cargo que el empleado tiene en la compañía.
- Proceso que apoya en la organización. Proceso en que se encuentra el empleado.
- Como valora la calidad de la información que utiliza en su proceso para las actividades que desarrolla. El objetivo de la pregunta es identificar el valor de calidad de la información para el empleado teniendo en cuenta que las posibles respuestas que se podrían responder son: Extremadamente satisfecho, Muy satisfecho, Moderadamente satisfecho, Poco satisfecho, no satisfecho.
- Frecuencia con la que se han realizado ejercicios para mejorar la calidad de los datos que utiliza para sus actividades. Con esta pregunta se pretende identificar si los empleados han realizado procesos de calidad y con qué frecuencia donde las opciones que se tienen son: Frecuentemente 3 veces al año, Algunas veces 1 vez al año, Pocas veces 1 vez cada dos años, Se hizo una vez, Nunca se ha hecho.
- Se publican las dimensiones referenciadas por los autores con el fin de que los usuarios encuestados puedan calificar que tan importante considera estas dimensiones para evaluar la calidad de la información en la empresa. Las dimensiones evaluadas fueron: Completitud, Consistencia, Precisión, Duplicidad, Validez, Antigüedad, Relevancia, Trazabilidad, Disponibilidad, Usabilidad, Accesibilidad, Interoperabilidad, Confiabilidad, Oportunidad. Para lo anterior se realiza una descripción general de las dimensiones que se evalúan como son:
    - Completitud: Se refiere a que los datos deben estar completos, en el sentido que en algunos casos la ausencia de éstos puede ser irrelevante, pero

cuando éstos se vuelven necesarios para un proceso esta información debe estar diligenciada.

- Consistencia - Integridad: Es el grado en que los datos almacenados en múltiples lugares son conceptualmente iguales.
- Precisión: Una medida de la exactitud del contenido de los datos. La precisión de los datos requiere la comparación de los datos con el objeto del mundo real que representan
- Duplicidad: Hace referencia a los datos que se repiten.
- Validez: Es el grado en que los valores del dato están en los rangos, límites, dominios y referentes esperados.
- Actualidad-Vigencia: Es la medida del grado en que los datos están actualizados en función del tiempo en que son consultados.
- Relevancia: Es el grado de importancia del dato
- Trazabilidad: Se puede hacer seguimiento al dato que se está trabajando y sus cambios en el tiempo.
- Disponibilidad: Que tan disponibles están los datos para las personas externas
- Usabilidad: Permite medir el grado de uso de estos datos en la empresa
- Accesibilidad: Se refiere al nivel de acceso de usuarios a los datos, si todos tienen acceso o algunos pocos, etc.
- Interoperabilidad: Se refiere a la capacidad, conocimiento y acuerdo de dos o más partes de un todo para interoperar.
- Confiabilidad: Se refiere a valorar que tan confiable son sus datos para la toma de decisiones en su empresa
- Oportunidad: Medida que permite conocer si el dato esa disponible cuando se requiere

Los encuestados en la empresa Central Hidroeléctrica de Caldas, hacen uso de datos y gestionan los datos generando reportes para usuarios finales o consolidando y diseñando soluciones que faciliten la gestión de los datos para los usuarios como son los analistas

de sistemas. La mayoría de los usuarios encuestados tienen conocimiento de herramientas tecnológicas y conceptos de sistemas lo cual facilito la socialización y gestión de la encuesta.

De este ejercicio se identifican aspectos que son importantes para el usuario final como son: la cantidad de usuarios que hacen uso de los datos y su nivel de importancia respecto a las dimensiones de calidad y por otra parte, cuales dimensiones asociadas a calidad de datos considera que la empresa debería darles mayor prioridad. La Figura 3.3 recoge la evaluación de dimensiones con mirada de usuario final.

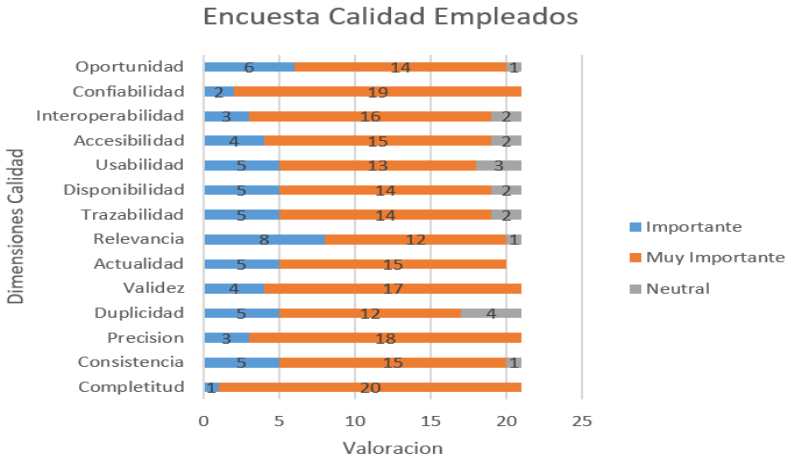


Figura 3.3 Calificación Dimensiones Empleados CHEC. Fuente: Elaboración Propia

Otra mirada complementaria al mapeo de dimensiones de calidad definida por los autores y la mirada de experto en el caso CHEC, es la satisfacción de la calidad de la información comparado con la frecuencia en la cual aplica calidad. Como se puede observar en la tabla 3.5 Análisis de Satisfacción de Calidad, se encuentra que un 57,1 por ciento de los encuestados manifiestan estar satisfechos de manera moderada con la calidad de la información; sin embargo, de este 100 por ciento de la población, el 50 por ciento solo aplico calidad una vez al año, un 41,7 por ciento ha aplicado calidad 3 veces al año y un 8,3 por ciento nunca ha aplicado calidad. En este punto, los datos indican que es poco el tiempo invertido en calidad de datos, aunque se encuentre moderadamente satisfecho el usuario.

Por otro lado, si se revisa la información asociada a la pregunta de frecuencia en la aplicación de calidad de datos, se concluye que de los usuarios que se encuentran muy satisfechos con la calidad de los datos, solo el 25,5 por ciento afirma haber aplicado calidad a sus datos 3 veces al año, otro 72,5 por ciento hacen calidad a sus datos en algún momento y un 12,5 por ciento no le han aplicado calidad a los datos.

Estos cuestionamientos son importantes a la luz de esta investigación, ya que surgen algunas preguntas como son:

- La calidad de datos no se aplica en la empresa porque no se conoce una metodología para aplicar?
- No se tienen las herramientas apropiadas para aplicar calidad de datos en la organización?
- El usuario no tiene como responsabilidad el garantizar la calidad de los datos que administra?

Teniendo en cuenta lo anterior, se concluye nuevamente que siendo la calidad de datos un tema relevante en las organizaciones, es una actividad que no está muy interiorizada en los procesos por lo cual el objetivo de la Tesis reviste mucha importancia para apalancar este tipo de necesidades.

Tabla 3.5 Análisis de Satisfacción Calidad Empleados CHEC. Fuente: Elaboración Propia

		%Satisf	%Frec
Extremadamente Satisfecho		4,8	
	Frecuentemente, 3 veces al año		4,8
Moderadamente Satisfecho		57,1	
	Algunas veces, 1 vez al año		50,0
	Frecuentemente, 3 veces al año		41,7
	Nunca se ha hecho		8,3
Muy Satisfecho		38,1	
	Algunas veces, 1 vez al año		25
	Frecuentemente, 3 veces al año		25
	Nunca se ha hecho		12,5

	Pocas veces, 1 vez cada 2 años		25
	Se hizo una vez		12,5

3.3 Selección de dimensiones

Posterior a estos ejercicios, se hizo una comparación de la relevancia de las dimensiones de calidad entre la revisión bibliográfica que fue revisada y el resultado de las encuestas aplicadas a los empleados de CHEC donde calificaron en orden de importancia las dimensiones, dicho resultado se puede observar en la Figura 3.4. donde se evidencia que algunas dimensiones tienen mayor importancia que otras, como se puede observar, dimensiones como Precisión, Consistencia, Disponibilidad, Completitud, Validez son las que están en el top de las preferencias por estos dos grupos de interés.

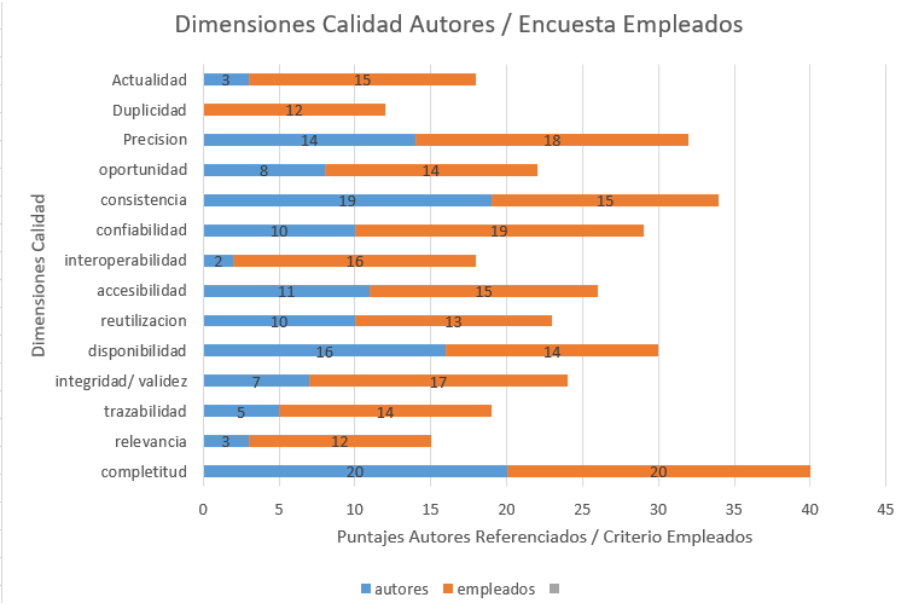


Figura 3.4 Asociación Dimensiones Calidad Autores / Encuesta Empleados CHEC Fuente: Elaboración propia

Con el fin de identificar las dimensiones más relevantes tanto para los autores en la revisión bibliográfica como para los empleados CHEC, se re-enumera en la columna Autores (de la tabla 3.5) el orden de importancia, el cual es asignado por la cantidad de

autores que referencian las dimensiones. Por otro lado, en la columna Empleados (de la tabla 3.5) se asigna una re-enumeración de dicha clasificación teniendo en cuenta la cantidad de empleados que marcaron como importante las dimensiones.

Finalmente, en la tabla 3.6 Mapeo dimensiones calidad en importancia Autores/Empleados, en la columna cuyo título es final, se asigna el puntaje que aparece dependiendo del promedio dado por los anteriores cfactores Autores y Empleados.

Tabla 3.6 Mapeo Dimensiones Calidad en importancia Autores/ Empleados Fuente: Elaboración Propia

Dimensiones mapeadas en importancia Autores/Empleados				
Dimension	**Final**	**Autores**	**Empleados**	**Referenciados**
Completitud	1	1	1	20
Confiabilidad	3	6	2	10
Consistencia	3	2	6	19
Accesibilidad	5	5	6	11
Precision	2	4	3	
Disponibilidad	4	3	7	16
Integridad/ validez	6	8	4	
Interoperabilidad		12	5	
Relevancia		11	9	
Actualidad		11	6	
Conformidad		10		
Comprensibilidad		10		
Trazabilidad		9	7	
Oportunidad		7	7	
Reutilizacion		6	8	
Duplicidad			9	

Del anterior ejercicio se toma la decisión que las dimensiones Completitud, Confiabilidad, Consistencia, Accesibilidad, Disponibilidad, Precisión e Integridad/Validez son las dimensiones que en un rango de 1 a 7 son las de mayor importancia en referencia por los autores y en la prioridad que le asignan los empleados en la encuesta previamente presentada. Es importante aclarar que este ejercicio no está descartando las otras dimensiones, simplemente se definen algunas para su analisis y que serán tenidas en cuenta para utilizar algunas de ellas en el prototipo del Modelo de Calidad de Datos que se propone en este estudio.

3.4 Dimensiones con métricas de autores

A partir de la revisión bibliográfica y de las dimensiones definidas en el anterior mapeo se identificaron algunas métricas referenciadas por autores y asociadas a cada dimensión, las cuales se describen a continuación.

La tabla 3.7 presenta por autor las métricas definidas para calcular la dimensión Completitud.

Tabla 3.7 Métricas Dimensión Completitud Autores Fuente: Elaboración propia

Dimension Completitud				
Autores				
(Ferney et al., 2018)	(Vetrò et al., 2016)	(Utamachant & Anutariya, 2018)	(Torchiano et al., 2017)	(Behkamal et al., 2014)
NCL = NR*NC: NCL numero de celdas, NR numero de filas NC número de columnas.	Porcentaje de celdas completas Porcentaje de filas completas	Celdas completas (CC) = Celdas incompletas / Total Celdas Filas Completas (CR) = Filas incompletas / Total Filas	PCC Porcentaje de celdas completas	M 19. Relacion de propiiedades a clases M 20. Perdida de informacion por caso
PPC = (1-(IC/NCL))*100: PPC Porcentaje de celdas competas, IC numero de celdas incompletas, NCL Numero de celdas calculadas con la formula previa.			PCRP Porcentaje de tuplas completas	
PCPR = (1-(NIR/NR))*100: PCPR Porcentaje de filas completas, NIR t numero de filas incompletas, NR numero de filas				

La tabla 3.8 presenta por autor la formulación definida por los mismos respecto a la dimensión de calidad Confiabilidad.

Tabla 3.8 Métricas Dimensión Confiabilidad Autores Fuente: Elaboración propia

Dimension Confiabilidad	
Autores	
(Ferney et al., 2018)	(Sadiq & Indulska, 2017)
PSC = (NS/NSC)*100: PSC Porcentaje de columnas estandarizadas NS Número de columnas con estándar NSC Número de columnas que comparten un estándar	Intención de comportamiento Expectativa de esfuerzo Condiciones facilitadoras Calidad de la información Expectativa de rendimiento Influencia social Calidad del sistema
eGMS = S + DC + C+ T: eGMS estándar que indica si los datos mas relevantes estan presentes. S fuente de datos DC fecha de creación C categoría T título	

La tabla 3.9 presenta por Dimensión de calidad Precisión, y autor las fórmulas definidas por los mismos.

Tabla 3.9 Métricas Dimensión Precisión/Integridad Autores Fuente: Elaboración propia

Dimension Precisión / Integridad				
Autores				
(Vetrò et al., 2016)	(Morbey, 2013)	(Torchiano et al., 2017)	(Behkamal et al., 2014)	(Utamachant & Anutariya, 2018)
Porcentaje de celdas acorde al dominio y tipo de información del dataset	¿Qué datos son incorrectos o caducados?	pcvc Porcentaje de celdas con valor correcto	M 1. Proporción de datos que contienen objetos faltantes M 2. Proporción de datos con objetos fuera de alcance M 3. Proporción de datos contiene un valor de datos mal escrito	Celdas precisas (SAC): [%] Porcentaje de celdas que tienen valores correctos. SAC = Celdas con valor correcto / Total celdas
				Celdas agregadas precisas (AAC): [%] Porcentaje de celdas que contienen validez AAC = Celdas correctas y precisas / Total celdas

La tabla 3.10 presenta por Dominio Consistencia y autor, las métricas definidas por los mismos para su aplicación.

Tabla 3.10 Métricas Dimensión Consistencia Autores Fuente: Elaboración propia

Dimension Consistencia			
Autores			
(Vetrò et al., 2016)	(Torchiano et al., 2017)	(Behkamal et al., 2014)	(Utamachant & Anutariya, 2018)
Porcentaje de columnas estandarizadas.	Duplicidad de datos en cada lote	M 16. Relación de datos que usan propiedades similares	Datos abiertos 5 estrellas (FS):
	Numero de atributos que se duplican en cada lote	M 17. Tipos de datos heterogenos	
		M 18. Valores inconsistentes	

Finalmente, con esta revisión bibliográfica se pudo identificar el interés de la comunidad en los problemas de calidad de datos y cuáles son las dimensiones más aplicadas y abordadas por los autores siendo esto el punto de referencia para iniciar el diseño del Modelo de Calidad de Datos aplicable para una empresa del sector como es CHEC.

A pesar de que se encontraron muchos trabajos alrededor del tema de datos abiertos, no existe un consenso en cuanto a las dimensiones relevantes que permitan determinar la calidad de los datos; no todos los autores publican las métricas asociadas a las dimensiones de calidad de datos que abordan en los artículos, la forma de calcularlos desde los datos abiertos y por otro lado se encuentra que muchos autores no presentan una definición de un índice de calidad de datos alineado con las dimensiones expuestas. Por todo lo anteriormente expuesto, esta Tesis se enfoca en proponer Modelo Conceptual que se concreta en un Modelo Tecnológico a través de un prototipo que permite calcular el Índice de Calidad de los Datos Abiertos para evaluar su calidad, basado en las métricas más representativas que arroja la revisión sistemática de artículos y los criterios de algunos usuarios de CHEC acerca de las dimensiones importantes basados en la experiencia de empresa.

4. Diseño del Modelo

Con el fin de enmarcar el resultado de esta Tesis, se entiende Modelo como un aporte conceptual donde se definen diferentes componentes (datos abiertos, dimensiones, métricas, índice, etc.) y la interrelación de estos para lograr el objetivo de obtener la evaluación de la Calidad de Datos Abiertos.

4.1 Dimensiones y Métricas

Basados en la revisión sistemática y el criterio de experto de CHEC en el anterior capitulo se definieron las dimensiones de mayor importancia las cuales son:

- **Completitud**

  Dimensión que permite controlar si un atributo definido para trabajar contiene sus datos completos o llenos. Es importante mencionar, desde la experiencia de negocio, no necesariamente el dato que este completo significa que este valido, por lo tanto, se hace relevante complementar este ejercicio con la dimensión validez de los datos para dar un valor más acertado al índice de calidad; sin embargo, este tipo de criterios deberán ser definidas por el usuario final quien es el dueño de los datos.

  En el Modelo propuesto se utiliza para esta dimensión, una de las métricas propuesta por el autor (Vetrò et al., 2016) cuya formulación corresponde a estas definiciones

  - Percentage of complete cells. Porcentaje de celdas completas
  - Percentage of complete rows. Porcentaje de filas completas

  En el Modelo se propone el cálculo de porcentaje de celdas completas, siendo esta la formulación que tiene mayor impacto ya que se revisan los datos a nivel de atributo.

  Para entender el proceso se explicará paso a paso la ejecución de este cálculo:

  - Total registros por atributo: Identificar cantidad de registros por cada atributo dentro de cada objeto, para lo cual se parte de los atributos que el

usuario final define deberán ser analizados. La tabla 4.1 presenta este cálculo en la columna Total Registros.

- Total registros ausentes: Identificar la cantidad de registros ausentes o faltantes por cada atributo dentro de cada objeto. Los atributos ausentes o faltantes es que en su contenido tengan un nulo. La tabla 4.1 visualiza este cálculo en la columna Total Registros Ausentes.
- Total registros completos: Identificar la diferencia de registros realmente completos la cual resulta de restar Cantidad de registros menos cantidad de registros ausentes. La tabla 4.1 presenta este cálculo en la columna Total Regs Completos.

Tabla 4.1 Ejemplo cálculo dimensión completitud para objeto Rayos Fuente: Elaboración propia

Atributo	Total Registros	Total Registros Ausentes	Total Regs Completos	%Completitud	%Atributo	%Atributo Calculado
circuito	1617	0	1617	100,00	14,30	14,30
codigo_departamento	1617	54	1563	96,70	14,30	13,80
codigo_municipio	1617	0	1617	100,00	14,30	14,30
codigo_subestacion	1617	54	1563	96,70	14,30	13,80
nombre_departamento	1617	0	1617	100,00	14,30	14,30
nombre_municipio	1617	0	1617	100,00	14,30	14,30
nombre_subestacion	1617	0	1617	100,00	14,20	14,20
					100,00	99,00

- Porcentaje Completitud: Para calcular el porcentaje de registros completos se debe dividir Total registros completos sobre total registros del objeto y luego multiplicarlo por cien. Este cálculo se realiza por cada atributo, en la tabla 4.1 se visualiza este cálculo en la columna %Completitud y su fórmula es la siguiente:

$$Porcentaje\ completitud\ = \left(\frac{Total\ Registros\ Completos}{Total\ registros}\right) * 100$$

Hasta el momento se ha explicado el cálculo de la dimensión completitud asociada a un objeto Rayos, y calculada de manera independiente por cada atributo; sin embargo, es necesario complementar dicho ejercicio con el cálculo real del atributo

dado el peso definido por el usuario final, para lo cual es importante explicar paso a paso el cálculo proyectado de la dimensión para todo el objeto:

- Definición de peso de los atributos: Este paso es necesario para el cálculo del Indicador de Calidad, ya que es necesario que el usuario final defina los atributos que requiere validar con el modelo y por cada uno de ellos deberá asignar un peso dependiendo de su importancia y su prioridad, esta actividad en adelante se le llamará parametrización de los atributos. El usuario final deberá garantizar que la suma de los porcentajes asignado a los atributos al final de un 100% como se muestra en la tabla 4.2 donde se visualiza cada atributo con su peso respectivo, para el ejemplo se colocaron los atributos con el mismo peso; sin embargo, esta definición puede variar.

Tabla 4.2 Ejemplo peso por atributos para el objeto Rayos definido por el usuario final. Fuente: Elaboración propia

Objeto	Atributos	Porcentaje
Rayos	Circuito	14,3
	Código_departamento	14,3
	Código_municipio	14,3
	Código _subestación	14,3
	Nombre_departamento	14,3
	Nombre_municipio	14,3
	Nombre_subestación	14,3
	Total Porcentaje	100

Los valores aquí definidos, se ven reflejados también en la tabla 4.1 en la columna %Atributo.

- Porcentaje Completitud por peso: Este cálculo es necesario para llevar el porcentaje de la dimensión completitud de cada atributo a su valor real dependiendo del peso real de cada atributo definido por el usuario. La formulación asociada requiere multiplicar el %Completitud por el %Atributo y dividirlo entre 100 para llevar dicho calculo a porcentaje. En la tabla 4.1 se puede visualizar este cálculo por cada atributo en la columna %Atributo calculado, al final la sumatoria

de los valores calculados debe estar en un rango de 0 a 100 porciento. Su formulación es:

$$Porcentaje\ Atributo\ por\ peso\ = \left(\frac{\%Completitud * \%Atributo}{100}\right)$$

- **Validez**.

  Dimensión que permite determinar si los valores de un atributo se encuentran dentro de los rangos, o valores establecidos definidos por los usuarios conocedores de los datos. Esta dimensión requiere una definición de reglas de negocio por cada atributo que se considere intervenir, y para ser más acertado en el índice de validez de calidad con este atributo es necesario que el usuario funcional, el conocedor del negocio y de los datos a revisar, defina las reglas que se deben implementar en este ejercicio. Esta dimensión fue incluida en Modelo propuesto, como referencia para el ejercicio se trabaja con el autor (Torchiano et al., 2017) donde se aplica la siguiente formulación:

  - pcvc Percentage of cells with correct value (value belonging to the domain) Same. Porcentaje de celdas con el valor correcto (valor definido por un dominio o rango de datos)

  Para entender el cálculo de esta dimensión en el Modelo es importante explicar paso a paso el proceso para dicho calculo:

  - Definición de atributos para aplicar cálculos de dimensión validez. En este punto el usuario final debe identificar en el objeto o estructura a analizar cual serían los atributos de este objeto que deberán ser objeto de estudio. En la tabla 4.3 se presenta para el ejercicio que el objeto seleccionado es rayos, donde para cada atributo seleccionado se especifica el rango de validez de la información lo que se define en las columnas valor_minimo y valor_maximo y el peso en valor de cada atributo está definido en la columna %atributo el cual define la importancia de esta información para el usuario final. Es importante aclarar que la sumatoria de la columna %atributo debe ser 100 porciento ya que es el rango en % valido que se

debe aplicar en esta formula, para el ejemplo presentado los atributos tienen el mismo peso, sin embargo, esto puede variar y por ende los cálculos también cambiarían.

Tabla 4.3 Ejemplo parametrización dimensión validez para el objeto Rayos definido por el usuario final. Fuente: Elaboración propia

tabla	atributo	valor_minimo	valor_maximo	%atributo
rayos	codigo_departamento	0	99	50,00
rayos	codigo_municipio	10000	17777	50,00

- Total registros por atributo: Identificar cantidad de registros por cada atributo dentro de cada objeto, para lo cual se parte de los atributos que el usuario final define deberán ser analizados. La tabla 4.4 visualiza este cálculo en la columna Total Registros.
- Total registros Novalidos: Identificar la cantidad de registros no validos dentro del rango definido por el usuario por cada atributo dentro de cada objeto. La tabla 4.4 visualiza este cálculo en la columna Total Registros Novalidos.
- Total registros validos: Identificar la diferencia de registros realmente validos la cual resulta de restar Cantidad de registros menos cantidad de registros Novalidos. La tabla 4.4 visualiza este cálculo en la columna Total Regs Validos.

Tabla 4.4 Ejemplo validez calculado para el objeto Rayos definido por el usuario final. Fuente: Elaboración propia

Atributo	Total Registros	Total Registros Novalidos	Total Regs Validos	% Validez	%Atributo	%Atributo Calculado
codigo_departamento	1617	0	1617	100,00	50,00	50,00
codigo_municipio	1617	407	1210	74,80	50,00	37,40
					100,00	87,40

- Porcentaje Validez: Para calcular el porcentaje de registros validos se debe dividir Total registros validos sobre total registros del objeto y luego multiplicarlo por cien. Este cálculo se realiza por cada atributo, en la tabla 4.3 visualiza este cálculo en la columna %Validez y su fórmula es la siguiente:

$$Porcentaje\ validez\ = \left(\frac{Total\ Registros\ Validos}{Total\ registros}\right) * 100$$

Para complementar el ejemplo con el cálculo real del atributo dado el peso definido por el usuario final, es importante recordar la tabla 4.3 donde en la columna %atributo el usuario definió un porcentaje para cada atributo.

- Porcentaje Validez por peso: Este cálculo es necesario para llevar el porcentaje de la dimensión validez de cada atributo a su valor real dependiendo del peso real de cada atributo definido por el usuario. La formulación asociada requiere multiplicar el %Validez por el %Atributo y dividirlo entre 100 para llevar dicho calculo a porcentaje. En la tabla 4.4 se puede visualizar este cálculo por cada atributo en la columna %Atributo calculado, al final la sumatoria de los valores calculados debe estar en un rango de 0 a 100 porciento. Su formulación sería:

$$Porcentaje\ Atributo\ por\ peso\ = \left(\frac{\%Validez * \%Atributo}{100}\right)$$

En la propuesta esta dimensión incluye los atributos y el rango de valores válidos para este ejercicio, lo cual se evidencia en la tabla 4.5

Tabla 4.5 Parametrizaciones de atributos por tabla para la dimensión validez

tabla	atributo	valor_minimo	valor_maximo	%atributo
rayos	codigo_departamento	0	99	50,00
rayos	codigo_municipio	10000	17777	50,00

- **Disponibilidad**.

Dimensión que permite identificar que tan actualizados están los datos en un periodo de tiempo. Esta dimensión se puede medir mediante un porcentaje que refleja si los datos se encuentran actualizados en cuyo caso su valor corresponderá a 100% y 0% cuando los datos han perdido su vigencia. Nuevamente este es un criterio adicional que el usuario funcional (usuario que trabaja en una empresa y que es conocedor del negocio y dueño de los datos) debe considerar y dependerá de su relevancia en el cálculo del índice de calidad. Para el ejercicio del modelamiento se toma como referencia la formulación presentada por el autor (Vetrò et al., 2016) donde se indica:

- Delay in publication. Demora en la publicación.
- El modelamiento propuesto, calcula con base en la fecha de actualización de los datos o su publicación en el portal de datos abiertos, teniendo en cuenta si dichos datos fueron actualizados en el mes que se está procesando el cálculo de calidad. Adicionalmente se debe aclarar que el cálculo se realiza sobre el objeto completo independiente si solo se actualizaron algunos registros, se calcula la dimensión disponibilidad del objeto completo.
- Para poder realizar este cálculo, se requiere tener en una estructura u objeto a nivel de base de datos, el registro de las fechas en las cuales los datos a procesar fueron actualizados en el portal. Es importante aclarar que esta información no se puede extraer automáticamente del portal, debe ser diligenciada de manera manual para posteriormente procesar su cálculo en cada periodo.

En la tabla 4.5 se presenta un ejemplo del cálculo del porcentaje de la dimensión disponibilidad para la estructura Rayos. Para la buena ejecución de los cálculos se requiere tener los datos de la fecha en que se publican los datos y la fecha en que se genera el proceso. Para explicar mejor los cálculos propuestos en el Modelo, se procede a tomar como ejemplo la estructura Rayos y periodo 2021 mes agosto.

- Fecha_Cálculo: Columna que representa la fecha en que se ejecuta la métrica, como ejemplo se explicará el ejercicio utilizando la información del 30 agosto 2021.
- Tabla: Contiene el nombre del objeto al cual se le aplicará el cálculo de índice de calidad, para el ejercicio se trabajará como ejemplo la estructura rayos.
- Fecha_ActDatosAbiertos: Fecha en la que se hizo la última publicación de información de los datos de la estructura de Rayos en el portal de Datos Abiertos.
- Año: Año de ejecución del proceso de calidad de Rayos para la dimensión disponibilidad. Se calcula tomando el año de la columna Fecha_Calculo.
- Mes: Mes de ejecución del proceso de calidad de Rayos para la dimensión disponibilidad. Se calcula tomando el mes de la columna Fecha_Calculo.
- DiasDatosActualizados: Días de actualización de datos, esta columna se calcula restando la fecha de publicación de los datos que se visualiza en la columna Fecha_ActDatosAbiertos menos la fecha en que se efectúa el cálculo de la dimensión la cual se visualiza en la columna Fecha_Calculo, se debe tener en cuenta que si la diferencia en días es mayor a 30 significa que no hay datos publicados actualizados en el mes que se está haciendo este cálculo.
- %Disponibilidad: Porcentaje disponibilidad, se calcula dependiendo del atributo a revisar, para cada atributo se debe validar la columna DiasDatosActualizados si esta es mayor a 30 días (parámetro por defecto considerado por experto), el porcentaje disponibilidad es cero de lo contrario sería 100.

$$\%Disponibilidad = \left\{ Atributo\ Clave \mid Si\ DiasDatosActualizados > Parametro\ experto, \%Disponibilidad = 0\% \;\middle|\; En\ otro\ caso, \%Disponibilidad = 100\% \right\}$$

- En el ejemplo que se presenta en la tabla 4.6, existe para la estructura Rayos el mes de septiembre donde DiasDatosActualizados es 13 pues la diferencia entre el 30 de septiembre y el 17 de septiembre da 13 días y como este valor es por debajo de 30 días entonces el porcentaje de disponibilidad será el 100%.

  Los anteriores cálculos se realizan de manera automática cada vez que se corren los algoritmos propuestos en el Modelo e implementados en el prototipo.

Tabla 4.6 Ejemplo cálculos dimensión disponibilidad para tabla rayos. Fuente: Elaboración propia

tabla	Fecha_Calculo	Fecha_ActDatosAbiertos	ano	mes	DiasDatosActualizados	%Disponiblidad
rayos	30/05/2021	29/06/2020	2021	5	335	0
rayos	30/06/2021	29/06/2021	2021	6	1	100
rayos	30/07/2021	29/06/2021	2021	7	31	0
rayos	30/08/2021	29/06/2021	2021	8	62	0
rayos	30/09/2021	17/09/2021	2021	9	13	100
rayos	30/10/2021	4/10/2021	2021	10	26	100
rayos	5/11/2021	2/11/2021	2021	11	3	100

- **Consistencia**

  Esta dimensión facilita a los usuarios determinar la cantidad de registros duplicados en la información que se está procesando. Como se trabajó la dimensión en el Modelo, permite la parametrización de un conjunto de atributos donde la combinación de ellos facilita la identificación de los registros duplicados en esta parametrización. Para el prototipo implementado, se define que la formulación a trabajar es la planteada por el autor (Torchiano et al., 2017) cuya formulación es la siguiente:

  - Consistency Duplication Number of participant which are duplicated in each lot. Consistencia en el número de componentes que estan duplicados en cada lote de datos.

Para la implementación del Modelo respecto a esta dimensión, es importante aclarar los pasos que se realizan para el cálculo de esta dimensión:

- Definición de variables. En este punto el usuario define las variables que van a ser utilizadas en el cálculo, por lo tanto, en la tabla 4.7 se visualizan los atributos definidos por el usuario final, que al momento del ejemplo se definieron año, mes, día y circuito. Para el Modelo se utiliza la combinación de estas variables como el lineamiento para validar si existen duplicidad de datos en un periodo de tiempo.

Tabla 4.7 Parametrizaciones de atributos por tabla para la dimensión consistencia. Fuente: Elaboración propia

tabla	atributo
rayos	ano
rayos	mes
rayos	dia
rayos	circuito

- Total registros por atributo: Identificar cantidad de registros por cada atributo dentro de cada objeto, para lo cual se parte de los atributos que el usuario final define deberán ser analizados. La tabla 4.8 visualiza este cálculo en la columna TotalRegs.
- Total registros Inconsistencia: Identificar la cantidad de registros inconsistentes que corresponde a la cantidad de registros duplicados por periodo que cumplen con la combinación de atributos definida por el usuario final, dicho calculo debe realizarse por cada atributo dentro de cada objeto. La tabla 4.7 visualiza este cálculo en la columna TotalRegInconsistencia.
- Ano: Identifica el año del atributo FechaCalculo
- Mes: Identifica el mes del atributo FechaCalculo
- Total registros consistentes: Identificar la diferencia de registros realmente validos la cual resulta de restar TotalRegs menos TotalRegInconsistencia. La tabla 4.8 visualiza este cálculo en la columna TotalRegsConsistencia.

Tabla 4.8 Ejemplo consistencia calculado para el objeto Rayos definido por el usuario final. Fuente: Elaboración propia

tabla	FechaCalculo	TotalRegs	TotalRegInconsistencia	ano	mes	TotRegsConsistencia	%Consistencia
rayos	20/11/2021	1574	40	2021	4	1534	97,46
rayos	20/11/2021	1574	50	2021	5	1524	96,82
rayos	20/11/2021	1574	50	2021	6	1524	96,82
rayos	20/11/2021	1574	2	2021	7	1572	99,87
rayos	20/11/2021	1574	2	2021	8	1572	99,87
rayos	20/11/2021	1574	2	2021	9	1572	99,87
rayos	20/11/2021	1574	2	2021	10	1572	99,87

- Porcentaje Consistencia: Para calcular el porcentaje de registros consistentes se debe dividir Total registros validos sobre total registros del objeto y luego multiplicarlo por 100. Este cálculo se realiza por cada atributo, en la tabla 4.7 se muestra este cálculo en la columna %Consistencia y su fórmula es la siguiente:

$$Porcentaje\ consistencia\ = \left(\frac{Total\ Registros\ Consistencia}{Total\ registros}\right) * 100$$

- **Confiabilidad**

  Esta dimensión es definida por varios autores como la capacidad de adherirse a estándares, convenciones o regulaciones (Ferney et al., 2018), aunque esta dimensión no fue utilizada en el modelo de propuesta actual y se espera incluirla en trabajo futuro. Su formulación de acuerdo con el autor referenciado es:

  - PSC = (NS/NSC)*100: Porcentaje de columnas estandarizadas donde NS el número de columnas con un estándar asociado sobre el número de columnas que cumplen con el estándar.

Teniendo en cuenta esta definición se aclara que el Modelo que se diseña debe permitir la parametrización de las dimensiones que se van a abordar con la Tesis, y

que para efectos del prototipo que se plantea es válido trabajar con algunas de las dimensiones que aquí se proponen.

4.2 Índice de Calidad

El Índice de Calidad, es un componente fundamental del Modelo final que reúne todos los cálculos de porcentaje de dimensiones de calidad definidas por el usuario final y da un resultado la calidad de los datos aplicando las métricas de la propuesta, basado en la parametrización final de las dimensiones definidas por el usuario. Es una formula integrada que consolida el cálculo de calidad de los datos con base en las parametrizaciones previamente definidas.

Para entender cómo se realiza este cálculo, es importante aclarar varios aspectos que se requieren para este proceso como son:

- El índice de Calidad es un valor que finalmente debe estar entre el rango de 0% – 100%, ya que todas las posibles dimensiones que se pueden trabajar están en porcentaje de 0 – 100.
- El cálculo del índice depende de las dimensiones a valorar, lo cual es definido por el usuario final sin embargo, se requiere que al final de parametrizar estos porcentajes la sumatoria de los porcentajes de las dimensiones definidas por el usuario y el rango de las mismas este entre 0% y 100%.
- Un ejemplo para entender esta configuración se presenta a continuación donde se define que las dimensiones Completitud, Validez, Disponibilidad y Consistencia se utilizaran en la parametrización de los componentes del Modelo en la tabla rayos, y que el porcentaje de Completitud corresponda a 40, el porcentaje de Validez corresponda a 30, el porcentaje de Disponibilidad corresponda a 20, el porcentaje de Consistencia corresponda a 10, cumpliendo lo requerido del 100 % total. Lo cual se aprecia en la tabla 4.9 del presente documento.

- Como el cálculo de las dimensiones depende de los atributos que se deben evaluar, y a su vez cada atributo debe tener un porcentaje asociado deberá permitir esa parametrización de variables. Esta definición paso a paso se evidencia en el capítulo 4.1 donde por cada dimensión se explica cómo se calculan los porcentajes a nivel de atributos.

Tabla 4.9 Ejemplo de configuración de porcentajes sobre las dimensiones de Rayos. Fuente: Elaboración propia.

FechaParametrizacion	Tabla	Dimension	%DimensiónTabla
5/09/2021	rayos	Completitud	40
5/09/2021	rayos	Validez	30
28/10/2021	rayos	Disponibilidad	20
9/11/2021	rayos	Consistencia	10
			100

- Es importante recordar que dichos porcentajes y parametrizaciones deben ser definidos por el usuario final, que es conocedor de sus datos y la importancia que un atributo pueda tener más que otro.
- La formulación para el cálculo del Índice de Calidad es la siguiente:

$$IndiceCalidad = \sum_{k=1}^{n} (\%\ Dimensión\ Tabla)$$

n: Se define al número de dimensiones involucradas en el cálculo del Índice de Calidad

Dimensión: Se refiere a la dimensión o las dimensiones definidas por el usuario en el modelamiento para el cálculo del índice de calidad.

A continuación, se explicará cómo se calcula el índice de calidad paso a paso por cada dimensión:

- %Completitud Tabla
  - Porcentaje Completitud por peso: Cálculo de variable definida en el capítulo 4.1 donde se detalla paso a paso su cálculo.

- Porcentaje dimensión por tabla: Se debe revisar el valor parametrizado en la tabla 4.8 que para efectos del ejercicio el porcentaje de completitud es un 40%.
- %Completitud Tabla: Se debe multiplicar el valor del %Completitud por el % dimensión por tabla definido para el modelamiento final y luego dividirlo entre 100. Este cálculo se puede visualizar en la tabla 4.10

$$\%\ Completitud\ Tabla\ = \frac{(Porcentaje\ Completitud\ por\ peso * \%DimensionTabla\ )}{100}$$

Tabla 4.10 Ejemplo de Cálculo Completitud total de la tabla Rayos. Fuente: Elaboración propia.

Atributo	ano	mes	Dimension	%Atributo_Calculado	%Completitud_Tbl
circuito	2020	3	Completitud	14,3	5,72
codigo_departamento	2020	3	Completitud	13,8	5,52
codigo_municipio	2020	3	Completitud	14,3	5,72
codigo_subestacion	2020	3	Completitud	13,8	5,52
nombre_departamento	2020	3	Completitud	14,3	5,72
nombre_municipio	2020	3	Completitud	14,3	5,72
nombre_subestacion	2020	3	Completitud	14,3	5,72
				99,1	39,64

- %Validez Tabla
  - Porcentaje Validez por peso: Cálculo de variable definida en el capítulo 4.1 donde se detalla paso a paso su cálculo.
  - Porcentaje dimensión por tabla: Se debe revisar el valor parametrizado en la tabla 4.8 que para efectos del ejercicio el porcentaje de validez es un 30%.
  - %Validez Tabla: Se debe multiplicar el valor del %Validez por el % dimensión por tabla definido para el modelamiento final y luego dividirlo entre 100. Este cálculo se puede visualizar en la tabla 4.11

$$\%\ Validez\ Tabla\ = \frac{(Porcentaje\ Validez\ por\ peso * \%DimensionTabla\ )}{100}$$

Tabla 4.11 Ejemplo de cálculo Validez total de la tabla Rayos. Fuente: Elaboración propia.

atributo	ano	mes	Dimension	%Atributo_Calculado	%Validez_Tbl
codigo_departamento	2020	3	Validez	50	15
codigo_municipio	2020	3	Validez	37,4	11,22
	2020	3	Validez	87,4	26,22

- %Disponibilidad Tabla
  - Porcentaje Disponibilidad: Cálculo de variable definida en el capítulo 4.1 donde se detalla paso a paso su cálculo.
  - Porcentaje dimensión por tabla: Se debe revisar el valor parametrizado en la tabla 4.8 que para efectos del ejercicio el porcentaje de Disponibilidad es un 20%.
  - %Disponibilidad Tabla: Se debe multiplicar el valor del %Disponibilidad por el % dimensión por tabla definido para el modelamiento final y luego dividirlo entre 100. Este cálculo se puede visualizar en la tabla 4.12

$$\%\ Disponibilidad\ Tabla\ = \frac{(Porcentaje\ Disponibilidad\ *\ \%DimensionTabla\ )}{100}$$

Tabla 4.12 Ejemplo de cálculo Disponibilidad total de la tabla Rayos. Fuente: Elaboración propia.

ano	mes	Dimension	%Atributo_Calculado	%Disponibilidad_Tbl
2020	1	Disponibilidad	0	0
2020	2	Disponibilidad	0	0
2020	3	Disponibilidad	0	0
2020	4	Disponibilidad	100	20
2020	5	Disponibilidad	0	0
2020	6	Disponibilidad	100	20
2020	7	Disponibilidad	0	0
2020	8	Disponibilidad	0	0

- %Consistencia Tabla
- Porcentaje Consistencia: Calculo de variable definida en el capítulo 4.1 donde se detalla paso a paso su cálculo.

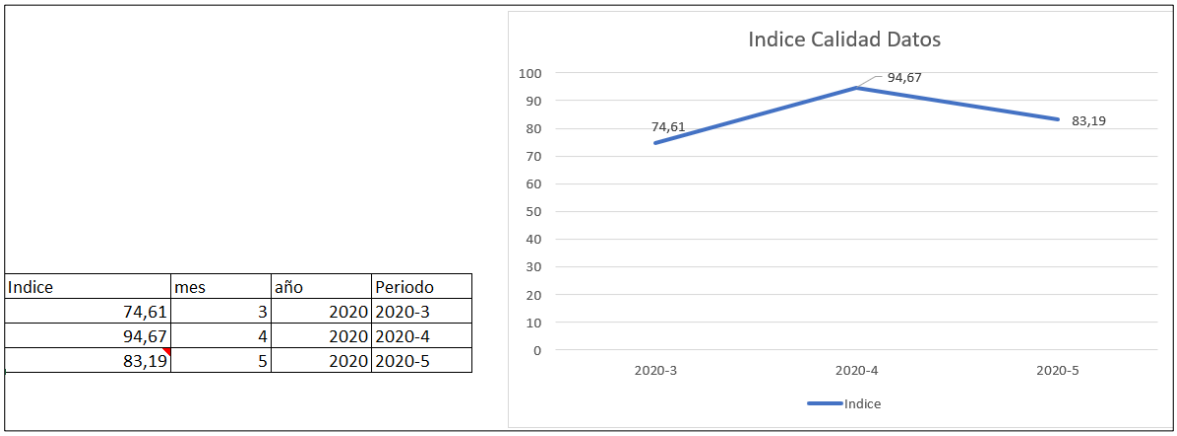
- Porcentaje dimensión por tabla: Se debe revisar el valor parametrizado en la tabla 4.8 que para efectos del ejercicio el porcentaje de Disponibilidad es un 10%.
- %Consistencia Tabla: Se debe multiplicar el valor del %Consistencia por el % dimensión por tabla definido para el modelamiento final y luego dividirlo entre 100. Este cálculo se puede visualizar en la tabla 4.13

$$\%\ Consistencia\ Tabla\ = \frac{(Porcentaje\ Consistencia\ * \%DimensionTabla\ )}{100}$$

Tabla 4.13 Ejemplo de cálculo Consistencia total de la tabla Rayos. Fuente: Elaboración propia.

ano	mes	Dimension	%Atributo_Calculado	%Consistencia_Tbl
2018	8	Consistencia	99,4	9,94
2018	10	Consistencia	99,8	9,98
2018	11	Consistencia	99,7	9,97
2019	2	Consistencia	99,9	9,99
2019	3	Consistencia	95,4	9,54
2019	5	Consistencia	99,3	9,93
2019	7	Consistencia	99,6	9,96
2019	9	Consistencia	98,4	9,84
2019	10	Consistencia	99	9,9
2019	11	Consistencia	99,8	9,98
2020	2	Consistencia	99	9,9
2020	3	Consistencia	87,5	8,75
2020	4	Consistencia	99,7	9,97

Una vez calculada cada una de las dimensiones con su respectivo porcentaje se realizan las sumatorias de las dimensiones parametrizadas hasta que finalmente se tiene un valor por mes del índice de calidad.



Indice	mes	año	Periodo
74,61	3	2020	2020-3
94,67	4	2020	2020-4
83,19	5	2020	2020-5

Figura 4.1 Ejemplo de cálculo Índice de calidad por meses de la tabla Rayos. Fuente: Elaboración propia.

4.3 Herramienta

Para la implementación del prototipo de solución se utilizó el motor de base de datos SQL Server donde el cálculo se realiza a nivel de bases de datos mediante un procedimiento almacenado en el servidor.

Por otro lado, en el proceso de consolidar los cálculos de calidad y trazabilidad de la calidad de los datos, se utilizó una herramienta de visualización como Power BI. Mediante la implementación de un dashboard se puede monitorear la calidad de la información en diferentes periodos del tiempo.

5. Validación de la solución

5.1 Fuente de Datos

La fuente de Datos Abiertos de CHEC para trabajar en esta Tesis es Rayos por Circuito que permite visualizar el total de Rayos por circuito de cobertura CHEC clasificados por año, mes, día cuyo contenedor de datos en el portal de MINTIC para datos abiertos es el "kscf-fk2u". La figura 5.1 permite evidenciar la ficha técnica del contenedor de datos.

Figura 5.1 Ficha técnica de contenedor Rayos por Circuito. Fuente: Portal MINTIC CHEC

En el ejercicio de implementar una solución que permita realizar una evaluación de la calidad de datos en CHEC, se diseña un Modelo Conceptual que se concreta en un artefacto tecnológico reflejado en el prototipo presentado, que permite almacenar de manera eficiente e incremental la información requerida para procesar las métricas que permiten determinar la calidad de los datos evaluados. La Figura 5.2 representa el modelo dimensional implementado para el prototipo.

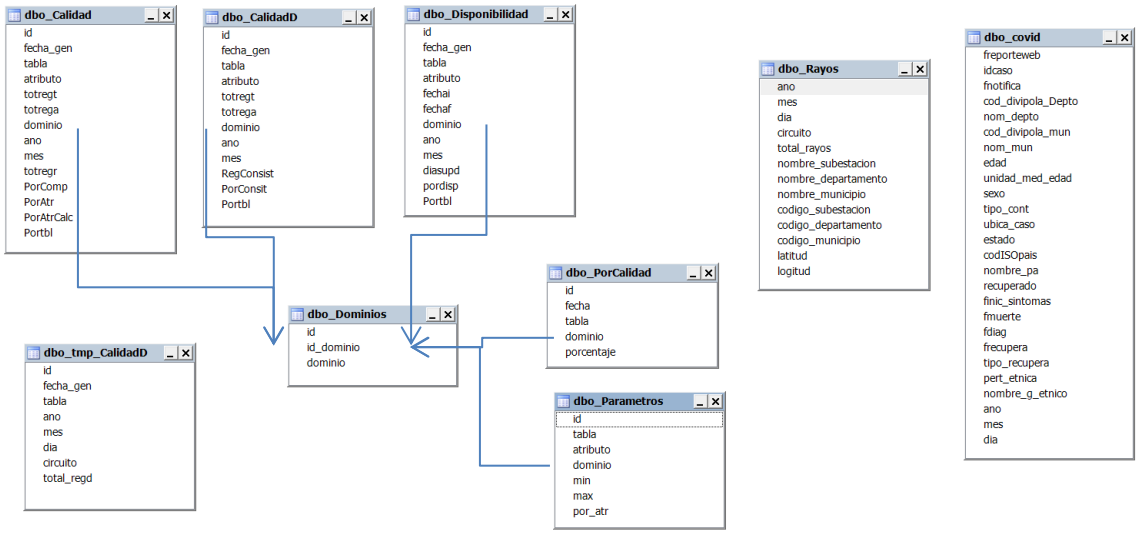


Figura 5.2 Modelo Dimensional para Calidad CHEC. Fuente: Elaboración Propia

Los objetos definidos en el modelo se describen a continuación:

- Dbo.Rayos. Estructura fuente que tiene la información del contenedor de Datos Abiertos.
- Dbo.Dominios. Objeto que permite parametrizar los dominios a evaluar en el modelo, para el ejercicio de la tesis se implementaron los dominios Completitud, Validez, Disponibilidad y Consistencia, teniendo claro que se pueden implementar más dimensiones, dicha información se muestra en la tabla 5.1.

Tabla 5.1 Tabla de dominios

id	id_dominio	dominio
1	1	Completitud
2	2	Validez
3	3	Disponibilidad
4	4	Consistencia

+

- Dbo.PorCalidad. Objeto que permite parametrizar el porcentaje de evaluación de cada dominio. Este porcentaje puede variar, lo importante es que la suma total de los porcentajes por dominio de 100, lo cual se evidencia en la figura 5.2.

Tabla 5.2 Tabla de porcentajes por dominio

id	fecha	tabla	dominio	porcentaje
1	5/09/2021	rayos	1	40
2	5/09/2021	rayos	2	30
5	28/10/2021	rayos	3	20
6	9/11/2021	rayos	4	10

- Dbo.Parametros. Objeto que permite configurar por tabla, atributo y dominio los parámetros válidos (atributos min-max) y los porcentajes que tendrá la valoración de cada atributo. Es importante aclarar que para un objeto la suma de sus atributos deberá ser el 100. Para el caso de dominio validez, se tienen dos variables a procesar código_departamento cuyo rango de validez esta entre 0-99 y para el atributo código_municipio el rango de validez está entre 10000 y 17777. En el caso de las variables asociadas al dominio validez, cada una se parametrizo con un 50 de valor que al sumarlos da un total de 100 porciento. Para el dominio Consistencia lo que se hace es configurar la combinación de variables que se requieren para que no se presenten datos duplicados. En este caso el porcentaje de atributo no es relevante porque la combinación de campos de un 100 por ciento. En la tabla 5.3 se evidencia esta parametrización.

Tabla 5.3 Tabla de porcentajes por atributo de una tabla.

id	tabla	atributo	dominio	min	max	por_atr
1	rayos	circuito	1	{null}	{null}	14,30
2	rayos	nombre_subestacion	1	{null}	{null}	14,30
3	rayos	nombre_departamento	1	{null}	{null}	14,30
4	rayos	nombre_municipio	1	{null}	{null}	14,30
5	rayos	codigo_subestacion	1	{null}	{null}	14,30
6	rayos	codigo_departamento	1	{null}	{null}	14,30
7	rayos	codigo_municipio	1	{null}	{null}	14,30
8	rayos	codigo_departamento	2	0	99	50,00
9	rayos	codigo_municipio	2	10000	17777	50,00
12	rayos	ano	4	{null}	{null}	{null}
13	rayos	mes	4	{null}	{null}	{null}
14	rayos	dia	4	{null}	{null}	{null}
15	rayos	circuito	4	{null}	{null}	{null}

- Dbo.Disponibilidad: Estructura que contiene el cálculo de la dimensión disponibilidad de los datos abiertos en la página MINTIC, donde dicha información

debe ser capturada de manera manual ya que no existe un mecanismo para extraer el histórico de esta captura de datos. En esta estructura aparte de almacenar la fecha de publicación de los datos, se calculan los días de actualización de la data, el porcentaje de disponibilidad de los datos y el porcentaje de cumplimiento del indicador, como ejemplo en la tabla 5.4 Ejemplo subconjunto de datos de Disponibilidad de rayos por circuito.

Tabla 5.4 Ejemplo subconjunto de datos Disponibilidad de Rayos por circuito

id	fecha_gen	tabla	fechaDato	dominio	ano	mes	diasupd	pordisp	Portbl
34	30/05/2020	rayos	20/04/2020	3	2020	5	40	0	0,00
25	30/06/2020	rayos	29/06/2020	3	2020	6	1	100	20,00
26	30/01/2020	rayos	29/06/2020	3	2020	7	151	0	0,00
27	30/08/2020	rayos	29/06/2020	3	2020	8	62	0	0,00
28	30/09/2020	rayos	29/06/2020	3	2020	9	93	0	0,00
29	30/11/2020	rayos	29/06/2020	3	2020	11	154	0	0,00
30	30/12/2020	rayos	29/06/2020	3	2020	12	184	0	0,00
580	31/03/2020	covid	2/03/2020	3	2020	3	29	100	15,00
581	30/04/2020	covid	11/03/2020	3	2020	4	50	0	0,00
582	31/05/2020	covid	24/03/2020	3	2020	5	68	0	0,00

- Dbo.CalidadD. Estructura donde se registran los cálculos del dominio Consistencia, para lo cual es necesario tomar como referencia los parámetros de consistencia definidos en la estructura dbo.Parametros. Un subconjunto de datos calculados para la dimensión Consistencia se puede visualizar en la tabla 5.5.

Tabla 5.5 Subconjunto de datos cálculo dimensión Consistencia de Rayos por circuito

id	fecha_gen	tabla	totregt	totrega	dominio	ano	mes	RegConsist	PorConsit	Portbl
8	8/11/2021	rayos	1874	2	4	2019	2	1872	99,89	4,99
9	8/11/2021	rayos	4540	208	4	2019	3	4332	95,42	4,77
10	8/11/2021	rayos	2762	18	4	2019	5	2744	99,35	4,97
11	8/11/2021	rayos	2937	12	4	2019	7	2925	99,59	4,98
12	8/11/2021	rayos	3063	50	4	2019	9	3013	98,37	4,92
13	8/11/2021	rayos	3625	38	4	2019	10	3587	98,95	4,95
14	8/11/2021	rayos	3186	6	4	2019	11	3180	99,81	4,99
15	8/11/2021	rayos	1042	10	4	2020	2	1032	99,04	4,95
16	8/11/2021	rayos	1617	202	4	2020	3	1415	87,51	4,38

- Dbo.Calidad. Objeto final donde el cálculo en el prototipo del modelo de calidad implementado, genera los resultados y estos son almacenados en esta estructura. Dicha estructura es de tipo incremental, lo que significa es que cada vez que el

prototipo sea ejecutado la información es almacenada en el repositorio de datos como se muestra en la tabla 5.6.

Tabla 5.6 Objeto que presenta el resultado de cálculo de calidad incluyendo porcentajes

id	fecha_gen	tabla	atributo	totregt	totrega	dominio	ano	mes	totregr	PorComp	PorAtr	PorAtrCalc	Portbl
3326	20/11/2021	rayos	circuito	260	0	1	2016	1	260	100,00	14,30	14,30	5,72
3327	20/11/2021	rayos	nombre_subestacion	260	0	1	2016	1	260	100,00	14,30	14,30	5,72
3328	20/11/2021	rayos	nombre_departamento	260	0	1	2016	1	260	100,00	14,30	14,30	5,72
3329	20/11/2021	rayos	nombre_municipio	260	0	1	2016	1	260	100,00	14,30	14,30	5,72
3330	20/11/2021	rayos	codigo_subestacion	260	0	1	2016	1	260	100,00	14,30	14,30	5,72
3331	20/11/2021	rayos	codigo_departamento	260	0	1	2016	1	260	100,00	14,30	14,30	5,72
3332	20/11/2021	rayos	codigo_municipio	260	0	1	2016	1	260	100,00	14,30	14,30	5,72
3333	20/11/2021	rayos	codigo_departamento	260	0	2	2016	1	260	100,00	50,00	50,00	15,00
3334	20/11/2021	rayos	codigo_municipio	260	65	2	2016	1	195	75,00	50,00	37,50	11,25

5.2 Monitoreo de Calidad

Una vez el prototipo se ejecuta, los datos son almacenados en las estructuras de consolidación de información que como fue mencionado en el anterior ítem corresponde a la estructura Dbo.Calidad, dbo.CalidadD, dbo.Disponibilidad. Sin embargo, como el usuario no interpreta esta información en tablas, se requiere implementar una interfaz visual al usuario; para lo cual se utilizó la herramienta Power BI donde se diseñó un dashboard que facilite la interpretación de estos cálculos.

En la Figura 5.3 se presenta el comportamiento del indicador de calidad en un periodo de tiempo (rango que el dashboard permite seleccionar en el filtro año y la estructura), lo cual facilita la visualización del comportamiento del indicador de calidad de datos en el tiempo. Se evidencia que el indicador tuvo un problema en su comportamiento en el mes de mayo ya que de un 94,65% que tenía en abril baja a 63,87% para el mes de mayo.

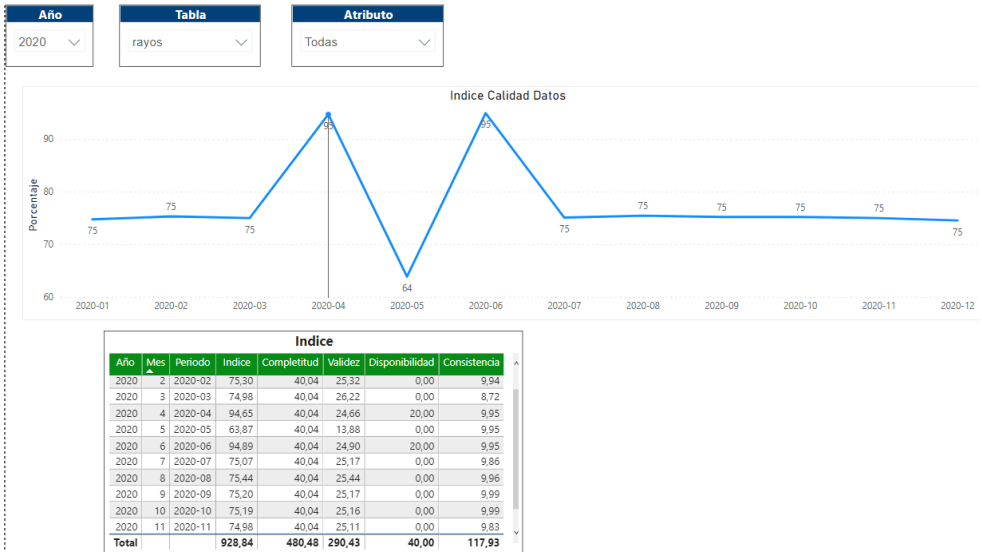


Figura 5.3 Historial Índice Calidad Datos Fuente: Elaboración propia

Para entender este comportamiento, se realiza otro reporte que permita al usuario entender con mayor detalle estos resultados, lo cual se visualiza en la Figura 5.4 donde se evidencia que las dimensiones que tuvieron problemas durante este período fueron validez y disponibilidad, en el atributo código_departamento donde hay 1657 registros de 2142 que no correspondían a datos válidos dentro del rango establecido para el código del departamento. Con estos resultados el usuario podrá aplicar de manera más oportuna los controles necesarios para garantizar que este problema no se vuelva a presentar.

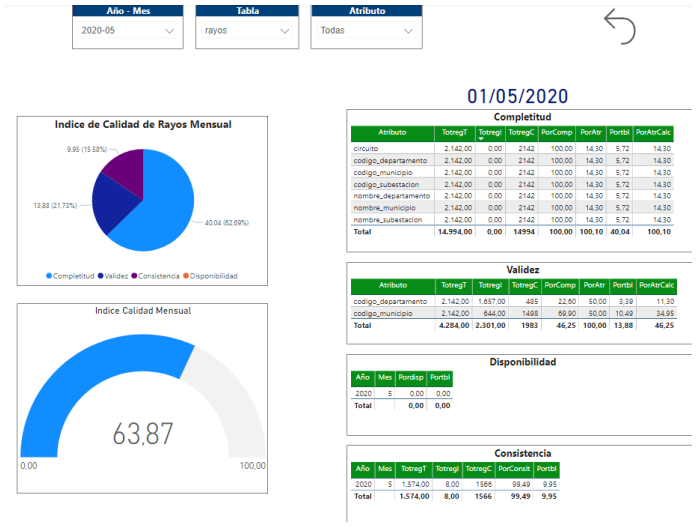


Figura 5.4 Calculo mensual del Índice de Calidad. Fuente: Elaboración propia

Por otro lado, se requiere visualizar de forma consolidada la información mensual de los cálculos de calidad con la evidencia de todos los atributos asociada a un objeto, para lo cual se construyó un reporte que contenga esta información. En la Figura 5.4 se visualiza un gráfico que se interpretan de la siguiente manera:

- Valores mensuales del índice: Consolidado que totaliza los valores de la Calidad de datos que se tiene durante el periodo elegido. Los valores se visualizan por dimensiones, iniciando con Completitud, luego con Validez, luego con Disponibilidad y por último Consistencia. Las columnas de las gráficas Completitud, Validez definen lo siguiente:
    - Atributo: Atributo a identificar
    - TotalregT: Cantidad de registros de la tabla
    - TotalRegI: Total registros incompletos
    - TotalRegC: Total registros completos

- PorCumpl: Calculo de % completitud
- PorAtr: Calculo de % de atributo completitud por el peso que tiene el campo en su parametrización
- PorCompTbl: Calculo de la completitud por tabla y detallado por atributo

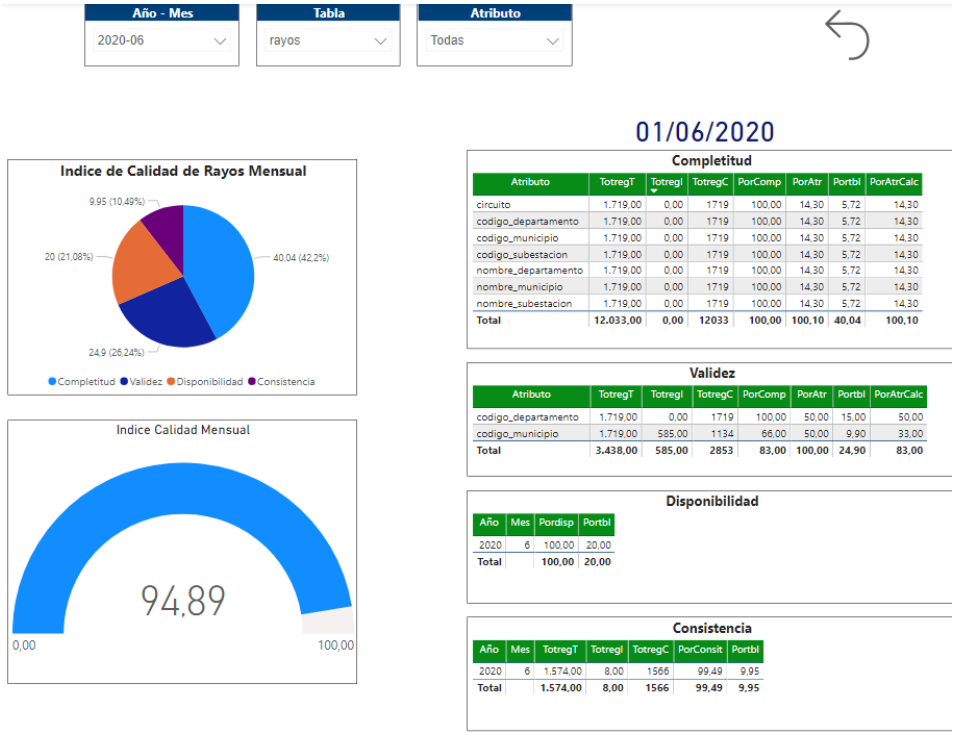


Figura 5.5 Calidad de Datos Mensual - Dimensiones. Fuente: Elaboración propia

- Índice de calidad de Rayos: Consolida por dimensiones el porcentaje calculado de la historia de los cálculos de calidad de datos a la fecha
- Validez: Consolidado que totaliza los valores de la historia de Calidad de datos que se tiene desde el primer periodo a la fecha. Los valores se visualizan por dimensiones, en el primer sector de la gráfica de arriba hacia abajo visualiza el consolidado de la dimensión validez. Las columnas de la gráfica definen lo siguiente:
  - Atributo: Atributo a identificar
  - TotalregT: Cantidad de registros de la tabla

  - TotalRegI: Total registros inválidos
  - TotalRegC: Total registros válidos
  - PorComp: Calculo dé % validez
  - PorAtr: Calculo dé % de atributo validez por el peso que tiene el campo en su parametrización
  - PorTbl: Calculo de la validez por tabla y detallado por atributo
- Disponibilidad: Dimensión que por cada año y mes indica:
  - Pordisp: Porcentaje disponibilidad calculado con base en la actualidad de los datos, si los datos están entre 1 y 30 dias actualizados el porcentaje será 100.
  - Portbl: Porcentaje que corresponde al equivalente en la parametrización de la dimensión.
- Consistencia: Dimensión que por cada año y mes indica:
  - TotalregT: Cantidad de registros de la tabla
  - TotalRegI: Total registros duplicados
  - TotalRegC: Total registros validos
  - PorConsist: Calculo dé % de atributo Consistencia por el peso que tiene el campo en su parametrización
  - PorTbl: Calculo de la consistencia por tabla y detallado por atributo

6. Conclusiones y recomendaciones

6.1 Conclusiones

El trabajo de Tesis permitió explorar, investigar y evaluar los artículos y propuestas que diferentes autores han investigado alrededor de la temática de Datos Abiertos. En este proceso se evidenciaron diferentes tendencias y las fortalezas y debilidades de las propuestas, lo que permitió encontrar un espacio para esta Tesis.

Desde el espacio encontrado en la revisión de la literatura y en el trabajo con los expertos esta Tesis presenta aporte en la definición de un Índice de Calidad de los Datos Abiertos compuesto métricas que permiten realizar los cálculos de diferentes dimensiones relacionadas con la calidad de datos, que pueden ser ponderadas según las condiciones particulares de cada fuente de datos, además de permitir adicionar nuevas dimensiones si fuese considerado, dada la modularidad de la propuesta.

Se cumple con el objetivo de determinar las características o dimensiones más relevantes que pueden ser vinculadas en el Modelo propuesto de Índice de Calidad de Datos, teniendo en cuenta que aparte de la mirada y la información compartida por los autores, se pudo complementar dicho trabajo con la perspectiva empresarial donde finalmente la empresa en la día a día evidencia problemas de calidad en los datos y este ejercicio se considera es enriquecedor para esta Tesis.

Como resultado se tiene un Modelo de Calidad de Datos que permite el cálculo de un Índice de Calidad que involucra diferentes dimensiones, que pueden ser ponderadas por los usuarios del sistema, y con métricas obtenidas en forma automática y semiautomática.

Al diseñar el prototipo de calidad de datos, se pudo cumplir con la expectativa de construir una herramienta basada en el Modelo propuesto que permita de manera automática o

semi-automática ejecutar una validación de calidad de datos en un pool de datos abiertos definidos por CHEC y que se puede aplicar a diferentes organizaciones.

6.2 Trabajo Futuro

Se considera importante y como complemento o trabajo a futuro la incorporación de nuevas dimensiones de calidad y el diseño de métricas correspondientes, que permitan robustecer esta propuesta que finalmente podría ser de mucha utilidad para las organizaciones que deben gestionar la calidad de datos.

Se pueden aprovechar los resultados para implementar técnicas de minería de datos que permitan encontrar patrones relacionados con la calidad de datos y para apoyar en la definición de las ponderaciones para el cálculo del Índice de Calidad.

Para el prototipo que se implementó en la Tesis, como cumplimiento del objetivo específico, se hizo uso de programación en el motor de base de datos SQL Server; sin embargo, no se alcanzó a implementar una plataforma completa de usuario final para facilitar la interacción con el mismo. Una de las mejoras es complementar la solución con una plataforma WEB que permita configurar las parametrizaciones y ejecutar de forma manual y cuando el usuario lo requiera este prototipo.

Bibliografía

25000, I. P. (2021). *Iso-25012 @ Iso25000.Com*. Iso 25000 Software and Data Quality. https://iso25000.com/index.php/en/iso-25000-standards/iso-25012

Abella, A., & De-pablos-heredero, M. O. C. (n.d.-a). *INDICADORES DE CALIDAD DE DATOS ABIERTOS : EL CASO DEL PORTAL DE DATOS ABIERTOS DE BARCELONA Open data quality metrics : Barcelona open data portal case*.

Abella, A., & De-pablos-heredero, M. O. C. (n.d.-b). INDICADORES DE CALIDAD DE DATOS ABIERTOS : EL CASO DEL PORTAL DE DATOS ABIERTOS DE BARCELONA Open data quality metrics : Barcelona open data portal case. *El Profesional de La Información*.

Abella, A., Ortiz-De-urbina-criado, M., & De-Pablos-heredero, C. (2019). Meloda 5: A metric to assess open data reusability. *Profesional de La Informacion*, *28*(6), 8–10. https://doi.org/10.3145/epi.2019.nov.20

Ahmed, H. H. (2018). Data quality assessment in the integration process of linked open data (LOD). *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, *2017-Octob*, 1–6. https://doi.org/10.1109/AICCSA.2017.178

Batini, C., & Scannapieca, M. (2006). *Data-Centric Systems and Applications: Data Quality Concepts, Methodologies and Techniques*.

Behkamal, B., Kahani, M., Bagheri, E., & Jeremic, Z. (2014). A metrics-driven approach for quality assessment of linked open data. *Journal of Theoretical and Applied Electronic Commerce Research*, *9*(2), 64–79. https://doi.org/10.4067/S0718-18762014000200006

Bicevskis, J., Bicevska, Z., Nikiforova, A., & Oditis, I. (2018). Data quality evaluation: a comparative analysis of company registers' open data in four European countries. *Computer Science and Information Systems*, *17*, 197–204. https://doi.org/10.15439/2018f92

Bonina, C., & Scrollini -Ilda, F. (n.d.). *Governing open health data in Latin America*.

Brezočnik, L., Fister, I., & Podgorelec, V. (2018). Swarm Intelligence Algorithms for Feature Selection: A Review. *Applied Sciences*, *8*(9), 1521. https://doi.org/10.3390/app8091521

Colborne, A., & Smit, M. (2018). Identifying and mitigating risks to the quality of open data in the post-truth era. *Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017*, *2018-Janua*, 2588–2594. https://doi.org/10.1109/BigData.2017.8258218

Congreso de la República. (2014). Ley 1712 de 2014. *Presidencia de La Republica*, 34. https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=56882

D´Agostino, M., Marti, M., Mejía, F., Cosio, G. de, & Faba, G. (2018). Estrategia para la gobernanza de datos abiertos de salud: un cambio de paradigma en los sistemas de información. *Revista Panamericana de Salud Pública*, *41*, e27. https://doi.org/10.26633/RPSP.2017.27

Daraio, C., Lenzerini, M., Leporelli, C., Naggar, P., Bonaccorsi, A., & Bartolucci, A. (2016). The advantages of an Ontology-Based Data Management approach: openness, interoperability and data quality. *Scientometrics*, *108*(1), 441–455. https://doi.org/10.1007/s11192-016-1913-6

Dawes, S. S., Vidiasova, L., & Parkhimovich, O. (2016). Planning and designing open government data programs: An ecosystem approach. *Government Information Quarterly*, *33*(1). https://doi.org/10.1016/j.giq.2016.01.003

Ferney, M. M. J., Beltran Nicolas Estefan, L., & Alexander, V. V. J. (2018). Assessing data quality in open data: A case study. *2017 Congreso Internacional de Innovacion y Tendencias En Ingenieria, CONIITI 2017 - Conference Proceedings*, *2018-Janua*, 1–5. https://doi.org/10.1109/CONIITI.2017.8273343

Giovannini, E. (n.d.). Towards a Quality Framework for Composite Indicators. *Oecd*.

Icontec. (n.d.). *Norma Tecnica Colombiana ISO 55001*.

ISO. (2021). *ISO/IEC 25012*. https://iso25000.com/index.php/en/iso-25000-standards/iso-25012

Kao, C. H., Hsieh, C. H., Chu, Y. F., Kuang, Y. T., & Yang, C. K. (2017). Using data visualization technique to detect sensitive information re-identification problem of

real open dataset. *Journal of Systems Architecture*, *80*(February), 85–91. https://doi.org/10.1016/j.sysarc.2017.09.009

Kubler, S., Robert, J., Neumaier, S., Umbrich, J., & Le Traon, Y. (2018). Comparison of metadata quality in open data portals using the Analytic Hierarchy Process. *Government Information Quarterly*, *35*(1), 13–29. https://doi.org/10.1016/j.giq.2017.11.003

Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC]. (2019a). *Ficha Tecnica Calidad de Datos*. https://herramientas.datos.gov.co/es/fichatecnicacalidad

Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC]. (2019b). *fichatecnicacalidad @ herramientas.datos.gov.co*. https://herramientas.datos.gov.co/es/fichatecnicacalidad

Ministerio de Tecnologías de la Información y Comunicaciones [MINTIC]. (2019c). *MAE . G . GEN . 01 – Documento Maestro*. 62. https://www.mintic.gov.co/arquitecturati/630/articles-144764_recurso_pdf.pdf

Mintic. (2011). *Manual de Gobierno Digital Implementación*. *2018*, 1–38.

Morbey, G. (2013). Data Quality for Decision Makers. In *Springer Gabler*. https://doi.org/10.1007/978-3-658-01823-8

Nikiforova, A. (2018). Open Data Quality Evaluation: A Comparative Analysis of Open Data in Latvia. *Baltic Journal of Modern Computing*, *6*(4), 363–386. https://doi.org/10.22364/bjmc.2018.6.4.04

Oliveira, J., Delgado, C., & Assaife, A. C. (2017). A recommendation approach for consuming linked open data. *Expert Systems with Applications*, *72*, 407–420. https://doi.org/10.1016/j.eswa.2016.10.037

PowerData. (n.d.). *Calidad de Datos. Cómo impulsar tu negocio con los datos.* Retrieved May 19, 2019, from https://www.powerdata.es/calidad-de-datos

PowerData. (2017). *¿Qué son los procesos ETL?* https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/qu-son-los-procesos-etl

PowerDAta. (n.d.). *Qué es un Data Lake y cómo funciona | Guía fácil y rápida*. Retrieved July 7, 2019, from https://www.mdirector.com/marketing-digital/data-

lake.html

Reiche, K. J., Höfig, E., & Schieferdecker, I. K. (2014). Assessment and visualization of metadata quality for government data. *CeDEM 2014, International Conference for E-Democracy and Open Government. Proceedings : 21-23 May 2014, Danube University Krems, Austria*, 335–346. http://publica.fraunhofer.de/documents/N-305654.html

Rengifo, S. C., Medina, L. F., & Tamayo, A. V. (2016). *Guía para el uso y aprovechamiento de Datos Abiertos en Colombia*.

Republica, presidencia de la. (2018). *Colombia, primer país en Latinoamérica con una política pública para la explotación de datos*. http://es.presidencia.gov.co/noticia/180417-Colombia-primer-pais-en-Latinoamerica-con-una-politica-publica-para-la-explotacion-de-datos

Ríos Ramírez, A., & Garro, J. E. (2018). Accountability y sociedad civil: el control político en la era digital. *Papel Político*, *22*(2), 311. https://doi.org/10.11144/javeriana.papo22-2.ascc

Ruijer, E., Grimmelikhuijsen, S., van den Berg, J., & Meijer, A. (2020). Open data work: understanding open data usage from a practice lens. *International Review of Administrative Sciences*, *86*(1), 3–19. https://doi.org/10.1177/0020852317753068

Sadiq, S., & Indulska, M. (2017). Open data: Quality over quantity. *International Journal of Information Management*, *37*(3), 150–154. https://doi.org/10.1016/j.ijinfomgt.2017.01.003

Safarov, I., Meijer, A., & Grimmelikhuijsen, S. (2017). Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Utrecht School of Governance*, *22*(1), 1–24. https://doi.org/10.3233/IP-160012

Santos, P. X. dos, & Guanaes, P. (2018). Ciência aberta, dados abertos: desafio e oportunidade. *Trabalho, Educação e Saúde*. https://doi.org/10.1590/1981-7746-sol00120

Scavuzzo, M., Nitto, E. Di, & Ardagna, D. (2018). Experiences and challenges in building a data intensive system for data migration. *Empirical Software Engineering*. https://doi.org/10.1007/s10664-017-9503-7

Talukder, M. S., Shen, L., Hossain Talukder, M. F., & Bao, Y. (2018). Determinants of user acceptance and use of open government data (OGD): An empirical investigation in Bangladesh. *Technology in Society*, *July*, 0–1. https://doi.org/10.1016/j.techsoc.2018.09.013

Torchiano, M., Vetro, A., & Iuliano, F. (2017). Preserving the Benefits of Open Government Data by Measuring and Improving Their Quality: An Empirical Study. *Proceedings - International Computer Software and Applications Conference*, *1*, 144–153. https://doi.org/10.1109/COMPSAC.2017.192

Utamachant, P., & Anutariya, C. (2018). An Analysis of High-Value Datasets: A Case Study of Thailand's Open Government Data. *Proceeding of 2018 15th International Joint Conference on Computer Science and Software Engineering, JCSSE 2018*, 1–6. https://doi.org/10.1109/JCSSE.2018.8457350

Veljković, N., Bogdanović-Dinić, S., & Stoimenov, L. (2014). Benchmarking open government: An open data perspective. *Government Information Quarterly*, *31*(2), 278–290. https://doi.org/10.1016/j.giq.2013.10.011

Verhulst, S. G., & Young, A. (2017). *OPEN DATA IN DEVELOPING ECONOMIES Toward Building an Evidence Base on What Works and How The GovLab*. www.odimpact.org

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016a). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, *33*(2), 325–337. https://doi.org/10.1016/j.giq.2016.02.001

Xia, W., Xu, Z., & Mao, C. (2018a). User-driven filtering and ranking of topical datasets based on overall data quality. *Proceedings - 2017 14th Web Information Systems and Applications Conference, WISA 2017*. https://doi.org/10.1109/WISA.2017.24

Xia, W., Xu, Z., & Mao, C. (2018b). User-driven filtering and ranking of topical datasets based on overall data quality. *Proceedings - 2017 14th Web Information Systems and Applications Conference, WISA 2017*, *2018-Janua*(1), 257–262. https://doi.org/10.1109/WISA.2017.24

Yoon, S. P., Joo, M. H., & Kwon, H. Y. (2019). How to guarantee the right to use PSI in the age of open data: Lessons from the data policy of South Korea. *Information Polity*, *24*(2), 131–146. https://doi.org/10.3233/IP-180103

Zhang, P., Xiong, F., Gao, J., & Wang, J. (n.d.). *Data Quality in Big Data Processing: Issues, Solutions and Open Problems*.

Zhang, P., Xiong, F., Gao, J., & Wang, J. (2018). Data quality in big data processing: Issues, solutions and open problems. *2017 IEEE SmartWorld Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation, SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI 2017* - , 1–7. https://doi.org/10.1109/UIC-ATC.2017.8397554

Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering Data Quality Problems: The Case of Repurposed Data. *Business and Information Systems Engineering*, *61*(5), 575–593. https://doi.org/10.1007/s12599-019-00608-0

Zhu, Y., & Cai, L. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, *14*(2), 1–10. https://doi.org/http://doi.org/10.5334/dsj-2015-002

7. Anexos

7.1 Datos Abiertos Covid

7.1.1 Fuente de datos

La fuente de Datos Abiertos adicional para trabajar en esta Tesis es Covid que permite visualizar el total de Rayos por circuito de Cobertura CHEC clasificados por año, mes, día cuyo contenedor de datos en el portal de MINTIC para datos abiertos es el "gt2j-8ykr". La figura 7.1 permite evidenciar la ficha técnica del contenedor de datos.

Figura 7.1 Ficha técnica de contenedor Covid. Fuente: Portal MINTIC CHEC

Los objetos definidos en el prototipo que se actualizan con la estructura Covid, se describen a continuación:

- Dbo.PorCalidad. Objeto que permite parametrizar el porcentaje de evaluación de cada dominio. Este porcentaje puede variar, lo importante es que la suma total de los porcentajes por dominio de 100, lo cual se evidencia en la figura 5.2.

Tabla 7.1 Tabla de porcentajes por dominio

id	fecha	tabla	dominio	porcentaje
3	18/09/2021	covid	1	30
4	18/09/2021	covid	2	50
7	15/11/2021	covid	3	15
8	15/11/2021	covid	4	5

- Dbo.Parametros. Objeto que permite configurar por tabla, atributo y dominio los parámetros validos (atributos min-max) y los porcentajes que tendrá la valoración de cada atributo. Es importante aclarar que para un objeto la suma de sus atributos deberá ser el 100. Para el caso de dominio validez, se tiene la variable cod_divipola_depto cuyo rango de validez esta entre 0-99. En este caso el porcentaje de atributo no es relevante porque la combinación de campos de un 100 por ciento. En la tabla 7.2 se evidencia esta parametrización.

Tabla 7.2 Tabla de porcentajes por atributo de una tabla.

id	tabla	atributo	dominio	min	max	por_atr
11	covid	cod_divipola_depto	1	{null}	{null}	50,00
16	covid	tipo_recupera	1	{null}	{null}	30,00
17	covid	nombre_g_etnico	1	{null}	{null}	20,00
10	covid	cod_divipola_depto	2	0	99	100,00

- Dbo.Disponibilidad: Estructura que contiene el cálculo de la dimensión disponibilidad de los datos abiertos en la página MINTIC, donde dicha información debe ser capturada de manera manual ya que no existe un mecanismo para extraer el histórico de esta captura de datos. En esta estructura aparte de almacenar la fecha de publicación de los datos, se calculan los días de actualización de la data, el porcentaje de disponibilidad de los datos y el porcentaje de cumplimiento del indicador, como ejemplo en la tabla 7.3 Ejemplo subconjunto de datos de Disponibilidad de covid.

Tabla 7.3 Ejemplo subconjunto de datos Disponibilidad de Rayos por circuito

id	fecha_gen	tabla	atributo	fechai	fechaf	dominio	ano	mes	diasupd	pordisp	Portbl
580	31/03/2020	covid	{null}	2/03/2020	{null}	3	2020	3	29	100	15,00
581	30/04/2020	covid	{null}	11/03/2020	{null}	3	2020	4	50	0	0,00
582	31/05/2020	covid	{null}	24/03/2020	{null}	3	2020	5	68	0	0,00
583	30/06/2020	covid	{null}	16/03/2020	{null}	3	2020	6	106	0	0,00
584	31/07/2020	covid	{null}	21/03/2020	{null}	3	2020	7	132	0	0,00
585	31/08/2020	covid	{null}	6/04/2020	{null}	3	2020	8	147	0	0,00

- Dbo.Calidad. Objeto final donde el cálculo de prototipo implementado de calidad, genera los resultados y estos son almacenados en esta estructura. Dicha estructura es de tipo incremental, lo que significa es que cada vez que el prototipo sea ejecutado la información es almacenada en el repositorio de datos como se muestra en la tabla 7.4.

Tabla 7.4 Objeto que presenta el resultado de cálculo de calidad incluyendo porcentajes

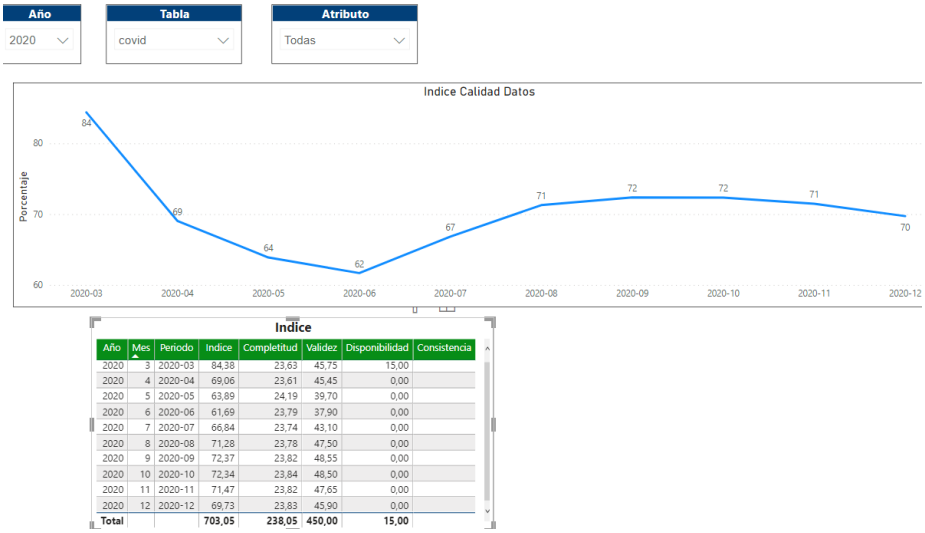
id	fecha_gen	tabla	atributo	totregt	totrega	dominio	ano	mes	totregr	PorComp	PorAtr	PorAtrCalc	Portbl
2620	15/11/2021	covid	cod_divipola_depto	905	0	1	2020	3	905	100,00	50,00	50,00	15,00
2621	15/11/2021	covid	tipo_recupera	905	44	1	2020	3	861	95,10	30,00	28,53	8,56
2622	15/11/2021	covid	nombre_g_etnico	905	895	1	2020	3	10	1,10	20,00	0,22	0,07
2623	15/11/2021	covid	cod_divipola_depto	905	77	2	2020	3	828	91,50	100,00	91,50	45,75

7.2 Monitoreo de Calidad

Una vez el prototipo se ejecuta, los datos son almacenados en las estructuras de consolidación de información que como fue mencionado en el anterior ítem corresponde a la estructura Dbo.Calidad, dbo.Disponibilidad. Sin embargo, como el usuario no interpreta esta información en tablas, se requiere implementar una interfaz visual al usuario; para lo cual se utilizó la herramienta Power BI donde se diseñó un dashboard que facilite la interpretación de estos cálculos.

En la Figura 7.2 se presenta el comportamiento del indicador de calidad en un periodo de tiempo (rango que el dashboard permite seleccionar en el filtro año y la estructura), lo cual facilita la visualización del comportamiento del indicador de calidad de datos en el tiempo. Se evidencia que el indicador tuvo un problema en su comportamiento en el mes

de junio ya que de un 63,89% que tenía en mayo baja a 61,69% para el mes de junio.



Año	Mes	Periodo	Indice	Completitud	Validez	Disponibilidad	Consistencia
2020	3	2020-03	84,38	23,63	45,75	15,00	
2020	4	2020-04	69,06	23,61	45,45	0,00	
2020	5	2020-05	63,89	24,19	39,70	0,00	
2020	6	2020-06	61,69	23,79	37,90	0,00	
2020	7	2020-07	66,84	23,74	43,10	0,00	
2020	8	2020-08	71,28	23,78	47,50	0,00	
2020	9	2020-09	72,37	23,82	48,55	0,00	
2020	10	2020-10	72,34	23,84	48,50	0,00	
2020	11	2020-11	71,47	23,82	47,65	0,00	
2020	12	2020-12	69,73	23,83	45,90	0,00	
Total			**703,05**	**238,05**	**450,00**	**15,00**	

Figura 7.2 Historial Índice Calidad Datos. Fuente: Elaboración propia

Para entender este comportamiento, se realiza otro reporte que permita al usuario entender con mayor detalle estos resultados, lo cual se visualiza en la Figura 7.3 donde se evidencia que las dimensiones que tuvieron problemas durante este período fueron completitud, validez y disponibilidad, en el atributo nombre_g_etnico se tienen resultados bajos por problemas de datos faltantes, y en la dimensión disponibilidad no hay datos actualizados en este periodo. Con estos resultados el usuario podrá aplicar de manera mas oportuna los controles necesarios para garantizar que este problema no se vuelva a presentar.

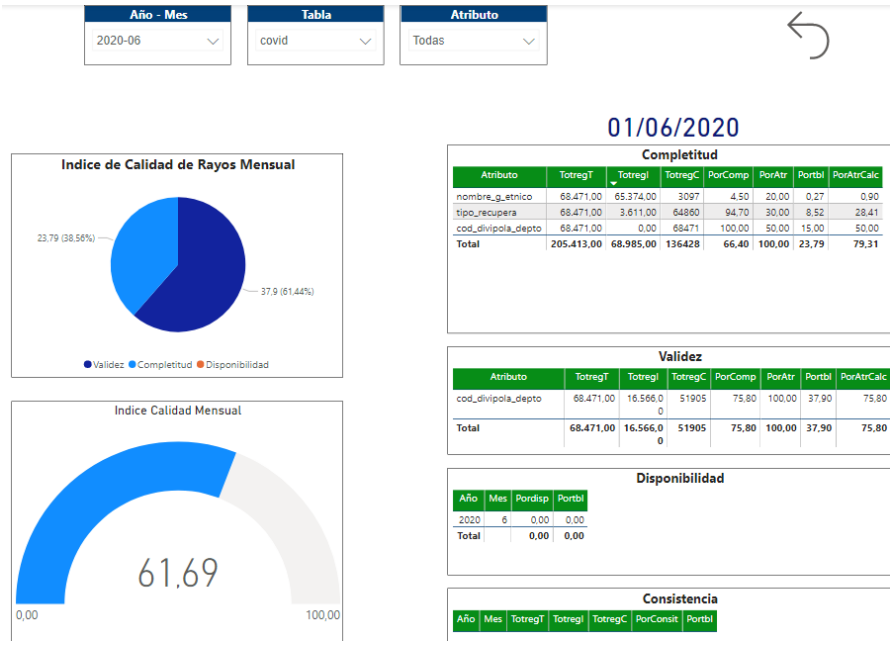


Figura 7.3 Calculo mensual del Índice de Calidad. Fuente: Elaboración propia

Por otro lado, se requiere visualizar de forma consolidada la información mensual de los cálculos de calidad con la evidencia de todos los atributos asociada a un objeto, para lo cual se construyó un reporte que contenga esta información. En la Figura 7.3 se visualiza un gráfico que se interpretan de la siguiente manera:

- Valores mensuales del índice: Consolidado que totaliza los valores de la Calidad de datos que se tiene durante el periodo elegido. Los valores se visualizan por dimensiones, iniciando con Completitud, luego con Validez, y Disponibilidad. Las columnas de las gráficas Completitud, Validez definen lo siguiente:
    - Atributo: Atributo a identificar
    - TotalregT: Cantidad de registros de la tabla
    - TotalRegI: Total registros incompletos
    - TotalRegC: Total registros completos

  - PorCumpl: Calculo de % completitud
  - PorAtr: Calculo de % de atributo completitud por el peso que tiene el campo en su parametrización
  - PorCompTbl: Calculo de la completitud por tabla y detallado por atributo

- Índice de calidad de Rayos: Consolida por dimensiones el porcentaje calculado de la historia de los cálculos de calidad de datos a la fecha
- Validez: Consolidado que totaliza los valores de la historia de Calidad de datos que se tiene desde el primer periodo a la fecha. Los valores se visualizan por dimensiones, en el primer sector de la gráfica de arriba hacia abajo visualiza el consolidado de la dimensión validez. Las columnas de la gráfica definen lo siguiente:
  - Atributo: Atributo a identificar
  - TotalregT: Cantidad de registros de la tabla
  - TotalRegI: Total registros invalidos
  - TotalRegC: Total registros validos
  - PorComp: Calculo dé % validez
  - PorAtr: Calculo dé % de atributo validez por el peso que tiene el campo en su parametrización
  - PorTbl: Calculo de la validez por tabla y detallado por atributo
- Disponibilidad: Dimensión que por cada año y mes indica:
  - Pordisp: Porcentaje disponibilidad calculado con base en la actualidad de los datos, si los datos están entre 1 y 30 dias actualizados el porcentaje será 100.
  - Portbl: Porcentaje que corresponde al equivalente en la parametrización de la dimensión.