

Gobernanza de Datos

Guía de **Calidad de Datos**

SECRETARÍA DE INNOVACIÓN Y TRANSFORMACIÓN DIGITAL

SUBSECRETARÍA DE POLÍTICAS PÚBLICAS BASADAS EN EVIDENCIA

Jefe de Gobierno

Horacio Rodríguez Larreta

Jefe de Gabinete

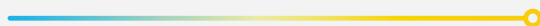
Felipe Miguel

Secretario de Innovación y Transformación Digital

Diego Fernández

Subsecretaria de Políticas Públicas Basadas en Evidencia

Melisa Breda



Índice

1. Introducción	2
2. Criterios generales de calidad	3
2.1. Criterios mínimos	3
2.2. Criterios básicos	4
2.3. Criterios óptimos	4
3. Criterios específicos	4
3.1 Criterios mínimos	5
3.1.1 Exactitud	5
3.1.1.1 Cualitativos	5
3.1.1.2 Cuantitativos	5
3.1.1.3. Pasos a seguir para verificar la exactitud:	6
3.1.2. Completitud	10
3.1.2.1. Cualitativos	10
3.1.2.2. Cuantitativos	10
3.1.2.3. Pasos a seguir para verificar la completitud:	11
3.1.2.4. ¿Cómo cuantificar los criterios mínimos?	12
3.2. Criterios Básicos	13
3.2.1. Consistencia	13
3.2.1.1. Cualitativos	13
3.1.1.2. Cuantitativos	14
3.2.1.3. Pasos a seguir para verificar la consistencia	15
3.2.2. Credibilidad	18
3.2.2.1. Cualitativos	18
3.2.2.2. Cuantitativos	18
3.2.2.3. Pasos a seguir para verificar la credibilidad	19
3.3.3. Actualidad	21
3.3.3.1. Cualitativos	21
3.2.2.4. ¿Cómo cuantificar los criterios básicos?	23
3.1. Criterios óptimos	24
3.3.1. Pertinencia	24
3.3.1.1. Cualitativos	24
3.3.1.2. Cuantitativos	24
3.3.1.3. Pasos a seguir:	25
3.3.2. Valor agregado	25
3.3.2.1. Cualitativos	25
3.3.2.2. Cuantitativos	25

3.3.2.3. Pasos a seguir	26
3.3.2.4. ¿Cómo cuantificar los criterios óptimos?	26
4. Score de calidad	27

En la era digital, los datos se han convertido en uno de los activos más valiosos para las organizaciones ya que desempeñan un rol fundamental en la gestión de procesos y toma de decisiones.

Reconociendo su importancia, el Gobierno de la Ciudad Autónoma de Buenos Aires (GCABA), a través de la Secretaría de Innovación y Transformación Digital (SECITD), ha adoptado una política de datos que busca fortalecer los procesos de obtención, integración y almacenamiento, con el objetivo de tomar decisiones basadas en evidencia y mejorar la vida de los vecinos que viven y transitan por la Ciudad.

Para lograr este propósito, el GCABA ha establecido una política de interoperabilidad como parte de su enfoque de gobernanza de datos. Estos lineamientos, basados en las mejores prácticas internacionales y el modelo DAMA-DMBOK, definen un marco común de gobernanza de datos que abarca todas las etapas del ciclo de vida de los mismos.

Con el objetivo de integrar y disponibilizar todos los datos del GCABA, en 2021 se implementó la Plataforma Inteligente de Buenos Aires (PIBA), a través de un data lake, que consiste en una herramienta de inteligencia aumentada.

En tanto, en 2022 se creó el Sistema de Interoperabilidad de la Ciudad de Buenos Aires (SI) con la intención de eficientizar los procesos y servicios de cara al ciudadano.

A continuación, se transmiten los lineamientos que rigen el data lake y que determinan los estándares que deben aplicarse en el ciclo de vida de gestión de datos. Los mismos establecen procesos claros y universales sobre cómo GCABA clasifica, comparte, accede, gestiona y protege los datos.

1. Introducción

Desde la Subsecretaría de Políticas Públicas Basadas en Evidencia se impulsa la cultura de datos en todo el Gobierno de la Ciudad Autónoma de Buenos Aires, siendo una de las responsabilidades primarias de este organismo, establecer lineamientos y estrategias en materia de administración y explotación de datos, actuando como responsable de implementar la política de datos y supervisar su cumplimiento.

En ese marco, la presente guía es uno de los lineamientos que tienden a establecer criterios unificados y concretar estándares en cuanto a la calidad de los datos. Es un documento práctico, que propone ser una herramienta de autoevaluación con el fin de ilustrar el estado, en materia de cumplimiento en los criterios de calidad, de la base de datos, previo a la utilización de ésta en la toma de decisiones.

Una de las políticas principales en materia de datos en el GCABA, es promover la reutilización de datos y construir datasets que puedan ser utilizados por todas las personas que así lo requieran, con el fin de generar valor tanto para organismos del sector público como instituciones del sector privado.

La toma de decisiones basadas en evidencias depende de que los datos cuenten con los criterios de calidad adecuados.

La presente guía se presenta como un compendio de directrices para mejorar la calidad de datos actuando directamente sobre cada uno de los criterios que la definen.

La estructura del documento comienza con una introducción a la calidad de datos, la definición de los criterios de calidad, con su división en: criterios mínimos, criterios básicos y criterios óptimos, con sus aspectos cualitativos y cuantitativos, finalizando, cada dimensión, con los pasos a seguir para su implementación.

2. Criterios generales de calidad



Fuente: Elaboración propia

Se considera que existen dos dimensiones que diferencian los criterios cualitativos de calidad de datos: la dimensión objetiva, que se relaciona directamente con el estado de la información; y la subjetiva, que apunta a optimizar la toma de decisiones.

Respecto de la dimensión objetiva, la fuente más frecuente de información para la medición de la calidad del producto son los resultados de las pruebas (estáticas o dinámicas) y, concretamente, los defectos localizados durante dichas pruebas.

La opción subjetiva suele estar asociada a la opinión del usuario, y/o la persona responsable técnica, es decir que suele basarse en encuestas y, necesitamos frecuentemente información antes de la puesta en producción.

2.1. Criterios mínimos

Los **criterios mínimos** comprenden exactitud y completitud, representa el nivel inicial en el proceso de medición de la calidad de los datos disponibles. Para ello, es necesario comprobar si los registros de las bases de datos **no se encuentran vacíos y si cumplen con las reglas de integridad**.

Esta evaluación se ejecuta únicamente para campos a nivel individual. La medición de calidad debe empezar por los criterios mínimos. Como parámetro general se puede indicar que un puntaje menor a 75% tanto en exactitud como en completitud tomados en conjunto, indicaría que se desestime el uso del campo analizado. No obstante, si bien podemos indicar a dicho 75% como un parámetro general, será responsabilidad de cada “Persona dueña técnica” definir los umbrales adecuados en función de la necesidad de cada caso de uso. A este parámetro lo llamaremos en adelante Umbral Límite Inferior Aceptable de Criterios Mínimos (ULIAM). En el caso de incumplimiento del ULIAM sería necesario reevaluar la selección de los campos o construir procedimientos para mejorar la calidad de los datos.

2.2. Criterios básicos

Los **criterios básicos** incluyen consistencia, credibilidad y actualidad. Tienen como finalidad generar confianza frente a la capacidad de evaluación de la calidad de los datos. De esta manera, se intenta verificar que los registros **sean congruentes entre sí en cada campo**; que se hayan sometido a un **proceso de validación**; y que los datos disponibles siempre se encuentren debidamente **actualizados**.

Como sucede con los criterios mínimos, esta evaluación se ejecuta únicamente para campos a nivel individual.

La medición de calidad deberá continuar por los criterios básicos. Como parámetro general se puede indicar que un puntaje menor a 75% tanto en consistencia, credibilidad y actualidad tomados en conjunto, indicaría que se desestime el uso del campo analizado. No obstante, si bien podemos indicar a dicho 75% como un parámetro general, será responsabilidad de cada “Persona dueña técnica” definir los umbrales adecuados en función de la necesidad de cada caso de uso.

Estos dos criterios (mínimos y básicos) se encuentran dentro de la **dimensión objetiva**, debido a que se relacionan intrínsecamente con el estado de la información, y por otro lado, son

medibles y comparables con umbrales preestablecidos, eliminando en ese sentido elementos de subjetividad.

2.3. Criterios óptimos

Los **criterios óptimos** indican que la organización cuenta con sistemas, procesos y prácticas de gobernanza de datos que informan la toma de decisiones basadas en evidencia.

Este criterio, a diferencia de los dos anteriores, conforma la **dimensión subjetiva**, en la medida que hace referencia a los criterios que orientan la capacidad estratégica de toma de decisiones, y se encuentran compuestos por la pertinencia y el valor agregado.

La medición de calidad deberá continuar por los criterios óptimos. Como parámetro general se puede indicar que un puntaje menor a 75% tanto en pertinencia, como en valor agregado, tomados en conjunto, indicaría que se desestime el uso del campo analizado. No obstante, si bien podemos indicar a dicho 75% como un parámetro general, será responsabilidad de cada “Persona dueña técnica” definir los umbrales adecuados en función de la necesidad de cada caso de uso.

3. Criterios específicos

Se presentan a continuación los Criterios Específicos, que son divididos, a su vez, en dos dimensiones, cualitativa y cuantitativa, con el objetivo de explicar en detalle cada una de sus características y atributos.

La dimensión cuantitativa se basa en variables medibles, utilizando datos de naturaleza numérica, como, por ejemplo, porcentaje y estadísticas.

La dimensión cualitativa está asociada a la idea de calidad, influenciada por el punto de vista del investigador. Se apoya en los datos obtenidos por el investigador de observaciones de primera mano, incluyendo cuestionarios, focus groups, observaciones participativas, grabaciones y documentos. Como regla general, los datos son no numéricos.

A continuación, se desarrollarán cada uno de los criterios específicos con su dimensión cualitativa y cuantitativa.

3.1 Criterios mínimos

Como se ha mencionado anteriormente, los criterios mínimos se encuentran comprendidos por los criterios de exactitud y completitud.

3.1.1 Exactitud

Se refiere a la diferencia entre el valor calculado y el valor real de la medición. Al trabajar con datos, es menester calcular la exactitud de las mediciones, con el fin de asegurar la producción de resultados válidos y objetivos.

Exactitud=1-error

El error es la diferencia entre el valor representado y el valor real.



3.1.1.1 Cualitativos

Representa el grado en el que los datos a completar en cada campo cumplen con la integridad indicada en un medio que los catalogue de forma unívoca y homogénea (por ejemplo: diccionarios). Esto implica que los datos están representados sin ambigüedad (mismas taxonomías), respetando símbolos, formatos y unidades.

La exactitud puede medirse en muchas dimensiones. Lo que significa específicamente, cómo se mide y qué resultado se considera aceptable, dependen siempre del caso de uso específico. Por ejemplo, en los archivos CSV, puede comprobarse la exactitud de cada celda de una columna con respecto a un formato de codificación para las fechas. La proporción entre celdas exactas e inexactas podría dar a las personas usuarias una primera impresión de lo que pueden esperar de los datos y de la dificultad de su procesamiento. Una mayor exactitud suele ser un indicador de datos de mayor calidad.

3.1.1.2 Cuantitativos

Cuando hablamos de exactitud en la dimensión cuantitativa, nos referimos a la diferencia entre el valor representado y el valor real, en cada campo.

Podemos decir que un campo es exacto o no en función del cumplimiento de un valor aceptable.

Al evaluar la conformidad de la columna "Tiempo de visualización" con la codificación ISO 8601, la tabla etiquetada como "mal ejemplo" obtendría una calificación de exactitud del 50 %, ya que la mitad de las celdas siguen este formato de tiempo. En cambio, la tabla etiquetada como "buen ejemplo" obtendría una puntuación de exactitud del 100 %, ya que todas las marcas de tiempo están correctamente codificadas.

Mal ejemplo	Buen ejemplo
Año; Visitantes; Tiempo de visualización	Año; Visitantes; Tiempo de visualización
2014;768954;3:18	2014;768954;00:03:18
2013;822101;00:02:59	2013;822101;00:02:59
2012;792697;0:02:52	2012;792697;00:02:52
2010;707402;3m:50s	2010;707402;00:03:50
2009;429430;3:16	2009;429430;00:03:16

3.1.1.3. Pasos a seguir para verificar la exactitud:

- **Paso 1:** identificar desde la metadata el tipo de dato del campo. Si el campo no hace parte en una base de datos relacional, entonces se debe verificar el tipo de registros disponibles para cada campo.

Variable	Tipo de dato BD
Tipo de documento	Numérico
Número de documento	Numérico
Nombre	Texto
Apellido	Texto
Teléfono	Alfanumérico
Género	Numérico

- **Paso 2:** establecer el tipo de dato esperado por campo.

Variable	Tipo de dato BD	Tipo de dato Esperado
Tipo de documento	Numérico	<p>✓ Se valida que es numérico porque corresponde a un listado de códigos, que está disponible en la base de datos.</p> <p>El listado de códigos contiene las opciones esperadas de acuerdo con el modelo de negocio del cliente.</p> <p>En este caso: DNI, LC, LE y documentos de identificación de extranjeros con un código específico por país.</p>
Número de documento	Numérico	✓ Numérico
Nombre	Texto	✓ Texto
Apellido	Texto	✓ Texto
Teléfono	Alfanumérico	✓ Numérico, no es el tipo de dato esperado, pero responde a la necesidad del modelo de negocio.
Género	Numérico	<p>✓ Se valida que es numérico porque corresponde a un listado de códigos, que está disponible en la base de datos.</p> <p>El listado de códigos contiene las opciones esperadas de acuerdo con el modelo de negocio del cliente.</p> <p>En este caso: Hombre, Mujer y no declara.</p>
Fecha de nacimiento	Numérico	✗ Se esperaba un campo tipo Fecha .
Fecha de alta	Fecha	✓ Fecha

- **Paso 3:** organizar o crear las reglas de integridad para cada campo de acuerdo con el tipo de dato esperado y con el objetivo de negocio del campo.

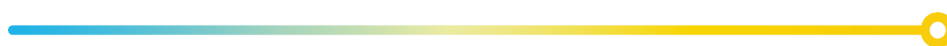
Tipo de documento	Los valores del campo están dentro del rango definido en el listado de códigos creado para el campo.
Número de documento	Para los números de documento que son DNI, el número está dentro del rango que "Base de datos 1" ha documentado en fuentes públicas. Para los tipos de documento para los que no es posible establecer un rango válido, se valida que sean campos numéricos.
Nombre Apellido	Regla 1. De acuerdo con la publicación de "Base de datos 1" se confirmó que la menor longitud posible es de 2 caracteres y la máxima de 25 caracteres. https://buenosaires.gob.ar/areas/registrocivil/nombres/busqueda/imprimir.php Se verifican los campos con longitudes más extremas, es decir, nombres/apellidos de 2 ó 3 caracteres y de más de 20 caracteres. Regla 2. Se detectan variaciones de 'NN' y se determinan cómo valores no válidos.
Variable	Condición de acuerdo al objetivo de negocio
Teléfono	Regla 1. Dado que no hay un parámetro específico para el ingreso de la información, se estima que la longitud posible del campo es de 8 a 12 caracteres. Únicamente para el análisis inicial se asume que un mismo número telefónico es usado solo por una persona. Por medio de una consulta se listan los valores únicos y la cantidad de veces que dicho valor ha sido registrado. Regla 2. Se detectan outliers que cumplen la regla 1 y que no son números válidos. Ejm. 11111111, 12345678, 22332233)
Género	Los valores del campo están dentro del rango definido en el listado de códigos creado para el campo.
Fecha de nacimiento	Por medio de información secundaria se determinó que la edad máxima puede ser de 120 años. El año actual se fija como el máximo posible y el año actual menos 120 como el mínimo posible. Ej: máximo 2022 y mínimo 1902.

Fecha de alta	Se determinó el día de hoy como la fecha máxima de ingreso y el 1 de enero de 2013 como la fecha mínima dado que el sistema se dió de alta durante ese año.
---------------	---

- **Paso 4:** evaluar que cada registro disponible en los campos cumpla con las reglas establecidas. Esto dará para cada registro un resultado binario, 1 será que cumple las reglas y 0 que no.
- **Paso 5:** se cuantifica la cantidad de registros que NO cumplen las reglas establecidas por campo. Se resta este valor al total de registros. Y para calcular el porcentaje de Exactitud se divide sobre el total de registros disponibles.

Los pasos 4 y 5 se muestran a continuación:

Variable	Total de Registros	Registros que NO cumplen la regla de integridad	% Exactitud por campo
Tipo de documento	334.067	0	100%
Número de documento	334.067	32.461	90.3%
Nombre	334.067	1.022	99.7%
Apellido	334.067	1.022	99.7%
Teléfono	334.067	28.180	91.6%
Género	334.067	0	100%
Fecha de nacimiento	334.067	62.587	81.3%
Fecha de alta	1.09.048	0	100%



3.1.2. Completitud

3.1.2.1. Cualitativos

Este criterio consiste en evaluar la cantidad de registros que no se encuentran vacíos.

A veces, los datos simplemente no están completos. Sin embargo, la falta de un valor no es razón para no utilizar o publicar los datos en cuestión. Para evitar confusiones, quien gestione los datos deberá marcar claramente los valores ausentes como valores nulos, ya que el valor nulo sirve como un marcador especial que indica que el valor no existe. En otras palabras, un valor nulo es una representación visual de un valor que falta. Hay varias formas de indicar un valor nulo, por ejemplo, marcando el valor que falta con "NULL" o "NA". Sin embargo, si se observa que dentro de sus datos tiene un alto porcentaje de valores nulos en una fila o columna, debería considerar la posibilidad de eliminar la columna o fila respectiva, ya que probablemente no aporte ningún valor añadido a los usuarios de los datos., o eventualmente se tomaría la decisión de completar los datos nulos con alguna de las técnicas típicamente utilizadas, como por ejemplo, la media o la mediana.

Ecuación matemática: La cantidad de valores no nulos sobre la cantidad de valores totales.

3.1.2.2. Cuantitativos

Cuando hablamos de completitud en su dimensión cualitativa, en el marco de la calidad de los datos, hacemos referencia al grado de disponibilidad de todos los datos de un conjunto de datos. De esta forma, una medida de la completitud de los datos es el porcentaje de campos que faltan.

Como mal ejemplo en algunos registros se encontraría nulos, como buen ejemplo, eliminar esos nulos, por la media o la mediana de visualización.

Mal ejemplo	Buen ejemplo
Año; Visitantes; Tiempo de visualización	Año; Visitantes; Tiempo de visualización
2014;768954;00:03:18	2014;768954;00:03:18
2013;;00:02:59	2013;null;00:02:59

2012;792697;00:02:52	2012;792697;00:02:52
2010;707402;	2010;707402;null
2009;429430;00:03:16	2009;429430;00:03:16

3.1.2.3. Pasos a seguir para verificar la completitud:

- **Paso 1:** conteo total de registros disponibles por campo, evaluando allí si el registro de un campo está supeditado a otro.
- **Paso 2:** conteo de registros nulos, vacíos y/o iguales a cero por campo.
- **Paso 3:** se resta el total de registros del paso 2 al total de registros del paso 1 y se divide sobre el total de registros del paso 1.

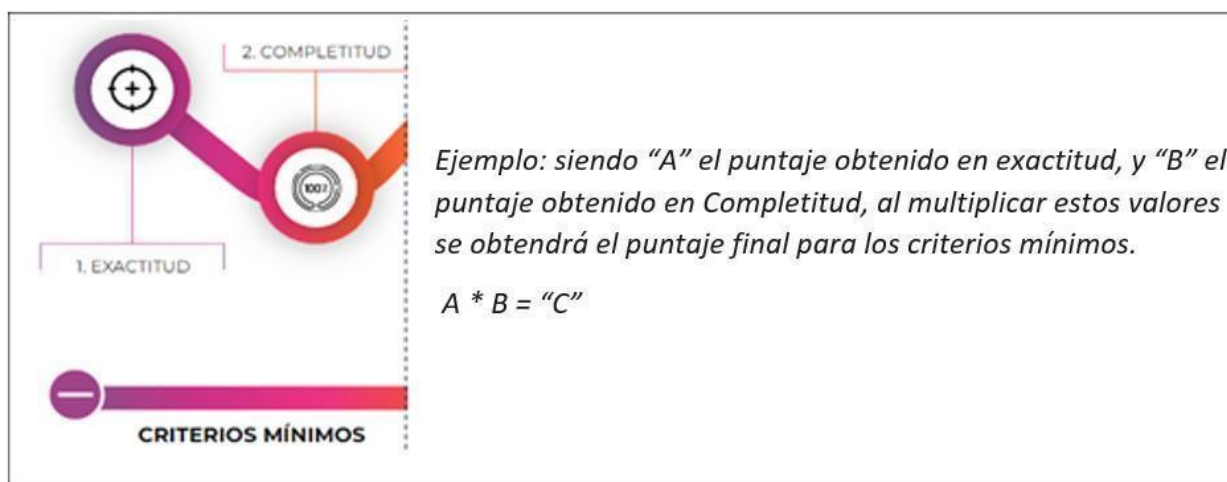
Los pasos 1, 2 y 3 se muestran a continuación:

Variable	Total de Registros	Registros Nulos, vacíos o 0s	% Completitud por campo
Tipo de documento	334.067	5.292	98.4%
Número de documento	334.067	1.400	99.6%
Nombre	334.067	0	100%
Apellido	334.067	0	100%
Teléfono	334.067	46.016	86.2%
Género	334.067	0	100%

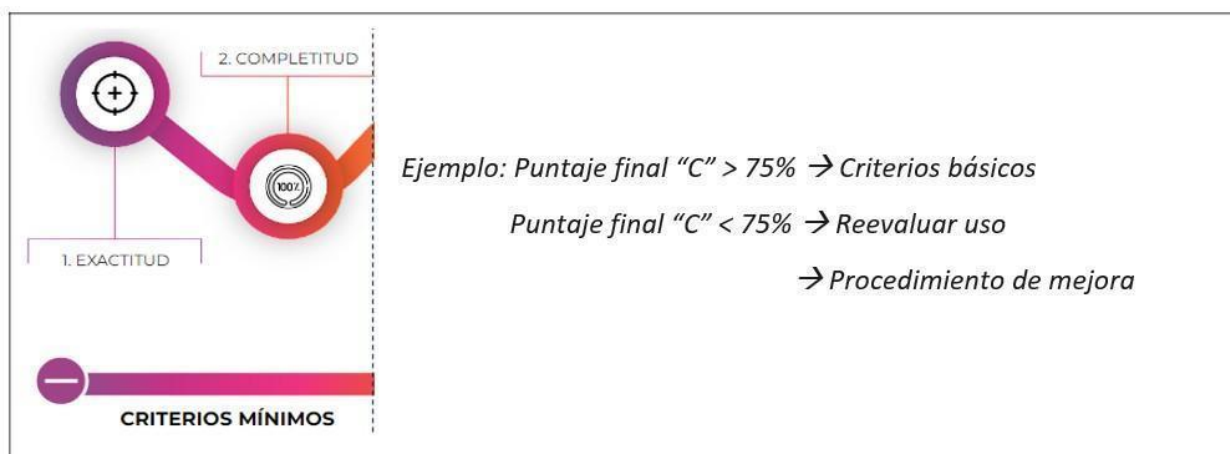
Fecha de nacimiento	334.067	19.607	94.1%
Fecha de alta	1.09.048	0	100%

3.1.2.4. ¿Cómo cuantificar los criterios mínimos?

1. Para cada campo individual se definirá un umbral límite inferior aceptable (ULIAM) de cumplimiento del criterio mínimo. Como parámetro general o en caso de indefinición se tomará una base del 75%. Este umbral deberá documentarse en el diccionario de metadatos.
2. Cada campo individual obtendrá un primer puntaje en los criterios mínimos. El campo entonces tendrá un puntaje en Exactitud y otro en Completitud, su multiplicación dará el puntaje final.



3. Para cada uno de los campos donde el puntaje de los criterios mínimos sea mayor o igual al ULIAM% se debería continuar con la medición de los criterios básicos.
4. Para aquellos campos con un puntaje total de los criterios mínimos menor al ULIAM será necesario reevaluar su uso o construir procedimientos para mejorar la calidad de los datos.



3.2. Criterios Básicos

3.2.1. Consistencia

3.2.1.1. Cualitativos

Esta medida representa la ausencia de diferencias entre los elementos de datos que representan los mismos objetos, los que no deberán tener contradicciones en sus bases de datos. Esto significa que, si se examinan dos valores de conjuntos de datos separados, coincidirán o se alinearán.

Los datos se pueden comparar por consistencia dentro de la misma base de datos (interna) o en comparación con otros conjuntos de datos de especificaciones similares (externa).

La consistencia medirá a nivel de cada registro la coherencia lógica entre campos de un mismo tipo de información (nominal, geográfica, datos de contacto, etc.) o por una regla de negocio. La evaluación se hará a nivel interno de la base entre los campos agrupados¹.

Todos los datos deberán respetar un mismo estándar, por ejemplo, la documentación deberá seguir una plantilla común o utilizar un vocabulario común.

Dependiendo del caso de uso, puede haber validadores que ayuden a comprobar los archivos con el estándar.

Garantizar la conformidad de los archivos con los estándares ayuda a la reutilización y facilita el procesamiento posterior. Para asegurarse de que sus datos se reutilicen, debería considerar el uso de estándares.

¹ El uso de la medición de consistencia es contingente, dado que la credibilidad puede reemplazar esta medición y mejorarla al evaluar el conjunto de datos en contraste con una base para la validación.

3.1.1.2. Cuantitativos

La consistencia en su dimensión cuantitativa se puede utilizar como una evaluación de la calidad de los datos y se puede medir como un porcentaje de los datos que reflejan el mismo tipo de información que la prevista para todo el conjunto de datos.

Por el contrario, los datos inconsistentes pueden incluir la presencia de atributos que no se esperan para la información prevista.

Por ejemplo, un conjunto de datos que contiene información sobre los usuarios de la aplicación se considera inconsistente si el recuento de usuarios activos es mayor que el número de usuarios registrados.

Mal ejemplo

Esta captura de pantalla muestra un mensaje de una validación SHACL que produjo un error contra el estándar utilizado. Más concretamente, el valor que se adjuntó a la propiedad `dcterms:publisher` no era del tipo requerido.

X	Lista de Servicios Web	
	<code>http://purl.org/dc/terms/publisher</code>	El valor no tiene clase <code>http://xmlns.com/foaf/01./Agent</code>

Buen ejemplo








Esta captura de pantalla muestra un conjunto de datos con un recurso XML que se ajusta a su esquema

O	Recursos	
	 DOWNLOAD	Archivo de sanciones 1.0 CSV
	 DOWNLOAD	Archivo de sanciones consolidadas 1.0 XML
	 DOWNLOAD	Archivo de sanciones consolidadas 1.1 CSV
	 DOWNLOAD	Archivo de Sanciones Consolidadas 1.1 XML
	Documentación	
	 DOWNLOAD	Archivo de Sanciones financieras consolidadas (XSD schema 1.0) XML SCHEMA

 DOWNLOAD	Archivo de Sanciones financieras consolidadas (XSD schema 1.1) XML SCHEMA
--	---

3.2.1.3. Pasos a seguir para verificar la consistencia

- **Paso 1:** el equipo técnico debe decidir operativamente cuáles campos serán agrupados por un mismo tipo de información o por una misma regla de negocio.

Variable	Tipo de Información o Regla de Negocio
Tipo de documento	 Datos de identificación
Número de documento	 Datos de identificación
Nombre	 Datos de identificación
Apellido	 Datos de identificación
Teléfono	 Datos de Contacto
Género	 Datos de identificación
Fecha de nacimiento	 Datos de identificación
Fecha de alta	-

- **Paso 2:** para la agrupación de campos por tipo de información:
 - **Paso 2.1.:** una vez agrupados los campos y medida la exactitud, se dispondrá de un resultado binario para cada registro (1 ó 0). Se deberá entonces sumar aquellos datos iguales a uno y se dividirá por el total de campos agrupados en el paso 1.

Variable Original	Variable Binaria
-------------------	------------------

ID Beneficiario	Tipo de documento	Número de documento	Nombre	Tipo de documento	Nro. de documento	Nombre	% Consistencia
1	1	2795263 1	NN	1	1	0	$\frac{2}{3} = 67\%$
2	1	2569874 1	Tomas	1	1	1	100%
3	1	222	Mariana	1	0	1	$\frac{2}{3} = 67\%$

- **Paso 2.2.:** se hará un promedio de los resultados por registro del paso 2.1 para obtener un puntaje total para los campos agrupados en el paso 1 para tipos de información.

Variable	% Consistencia por campo
Tipo de documento	78%
Número de documento	78%
Nombre	78%

- **Paso 3:** para la agrupación de campos por reglas de negocio:
 - **Paso 3.1.:** identificar los elementos que condicionan o determinan la relación establecida en las reglas de negocio.

Variable	Regla de Negocio
Fecha de nacimiento	Calcular la edad actual para cada beneficiario activo en un plan X.

- **Paso 3.2.:** se crea una consulta con los parámetros definidos en el paso 3.1 para cuantificar la cantidad de registros que cumplen la regla. La tabla ejemplifica la consulta:

ID	Nombre	Fecha de Nacimiento	Edad (Variable creada)	Descripción	Validación
1	NN	1/1/1900	(error)	Vivir en casa (Todos)	0
2	Tomas	28/04/1958	65.0	Vejez digna (Mayores de 63)	1
3	Mariana	06/06/2017	4.8	Escolaridad media (7 a 12 años)	0

- **Paso 3.3:** el resultado del paso 3.2 se divide en la totalidad de registros disponibles y allí se obtiene el puntaje de consistencia para cada agrupación realizada en el paso 1 para reglas de negocio.

Resultado: para el ejemplo el porcentaje de Consistencia de FK_Plan es del 33%, dado que solo 1 de los 3 registros analizados cumple la condición.

3.2.2. Credibilidad

3.2.2.1. Cualitativos

Este criterio trata de validar un conjunto de datos al contrastarlos con otra base que contenga tipos de informaciones asimilables.

Los datos se consideran creíbles si se basan en fuentes fiables. La credibilidad describe el grado en que los datos tienen atributos que los usuarios consideran verdaderos y creíbles

Por tanto, este indicador depende en gran medida de la percepción de los usuarios. No obstante, la credibilidad y la fiabilidad de los datos pueden aumentar si se proporciona cierta información contextual.

3.2.2.2. Cuantitativos

La credibilidad es la consistencia global de una medida. Se dice que una medida tiene una alta credibilidad si produce resultados similares en condiciones similares. Las puntuaciones que son altamente creíbles son precisas, reproducibles y consistentes de una prueba a otra. Es decir, si el proceso de evaluación se repitiera con un grupo de examinados, se obtendrían esencialmente los mismos resultados.

Ejemplo:

DOC BASE DE DATOS	DOC INTERNET
Cuentan con un proceso de revisión previo a la publicación.	No cuentan con mecanismo de control de calidad, y cualquier persona puede publicar en este medio sin tener experiencia o peritaje en el tema.
	<p>Para determinar la autoridad en Internet, se puede tomar en consideración los dominios o direcciones electrónicas.</p> <p>Las más recomendadas son: .edu (instituciones educativas). .gov (agencias del gobierno). .org (organizaciones). .mil (agencias militares).</p> <p>Las menos recomendadas son las que terminan en: .com (comercial) y (net)</p>

3.2.2.3. Pasos a seguir para verificar la credibilidad

- **Paso 1:** retomar la agrupación realizada en el criterio de consistencia. Se debe medir la credibilidad según los mismos criterios de agrupación.

Variable	Tipo de Información o Regla de Negocio
Tipo de documento	👤 Datos de identificación
Número de documento	👤 Datos de identificación
Nombre	👤 Datos de identificación
Apellido	👤 Datos de identificación
Teléfono	☎️ Datos de Contacto
Género	👤 Datos de identificación
Fecha de nacimiento	👤 Datos de identificación
Fecha de alta	-

- **Paso 2:** se determina cuál base de datos será útil para contrastar la agrupación de campos. Resulta recomendable que la base con la que se contraste se someta a criterios mínimos de calidad (exactitud y completitud) para garantizar mejores resultados.

Variable	Base de datos
Tipo de documento	👤 BASE DE DATOS 1 / SINTYS
Número de documento	👤 BASE DE DATOS 1 / SINTYS
Nombre	👤 BASE DE DATOS 1 / SINTYS
Apellido	👤 BASE DE DATOS 1 / SINTYS
Teléfono	☎️ BUKEALA / LOGIN
Género	👤 BASE DE DATOS 1 / SINTYS
Fecha de nacimiento	👤 BASE DE DATOS 1 / SINTYS
Fecha de alta	Variable generada por el sistema, no tiene verificación externa.

- **Paso 3:** normalizar los datos disponibles para poder someterlos a comparación con una base de datos externa.

En este ejemplo se evidencia que la cantidad de variables que usa es diferente, el campo se puede analizar concatenando todos los campos disponibles o llevando todos los campos a la unidad más atómica.

BASE DE DATOS 1		BASE DE DATOS 2		
Nombre	Apellido	NOMBRE	SEGUNDO NOMBRE	APELLIDO
Mario Heber	URQUIZA	MARIA		URQUIZA
Andrea	BOGADO	ANDREA		DUARTE
Constantino Salvatore	GALLO ROMERO	ORIANA	BELEN	GALLO ROMERO

- **Paso 4:** se crea una variable de chequeo binaria que expresa si el campo corresponde a la información de la base externa (=1) o si no coincide (=0).

BASE DE DATOS 2			
Nombre_1	Nombre_2	Apellido_1	Apellido_2
Mario	Heber	URQUIZA	
Andrea		BOGADO	
Constantino	Salvatore	GALLO	ROMERO

- **Paso 5:** se hará un promedio de los resultados por registro del paso 4 para obtener un puntaje total para los campos agrupados en el paso 1 para tipos de información.

Check				
NOMBRE	NOMBRE_2	APELLIDO_1	APELLIDO_2	% Credibilidad
0	-	1	-	50%
1	-	0	0	50%
0	0	1	1	50%

Resultado: para el ejemplo el porcentaje de Credibilidad de Nombre y Apellido es del 50%, ambas variables tendrían el mismo porcentaje.

3.3.3. Actualidad

3.3.3.1. Cualitativos

La actualidad dependerá de que el dato se encuentre vigente y en su última versión disponible, ya que existen registros que pueden someterse a modificaciones con diferente frecuencia (tanto para cambios eventuales, por ejemplo, residencia, teléfono, mail, etc, como para cambios recurrentes, por ejemplo, última fecha de cobro, edad, estado del beneficio).

Los metadatos y los datos son oportunos si están actualizados y representan la situación real y actual. Esto significa que en cuanto se produzca un cambio en el mundo real, los datos y los metadatos deberán modificarse también. Sin embargo, la evaluación de la actualidad no es tan trivial, ya que es difícil entender automáticamente a partir del contenido si se trata de datos históricos o en tiempo real. Por lo tanto, no es fácil determinar los requisitos de puntualidad de forma clara.

- **Paso 2:** contrastar la fecha de ingreso del registro² contra la regla de negocio establecida en el paso 1.

ID	Nombre	Fecha de Alta	Teléfono	Diferencia en años desde el registro a Hoy (Fecha Actual)	% Actualidad
1	NN	1/1/2014	2762-5631	(Año 2022 menos Año 2014) = 8 años	0%
2	Tomás	28/04/2021	3596-3682	(Año 2022 menos Año 2021) = 1 año	50%
3	Mariana	06/06/2022	8572-1464	(Año 2022 menos Año 2022) = 0 años	100%

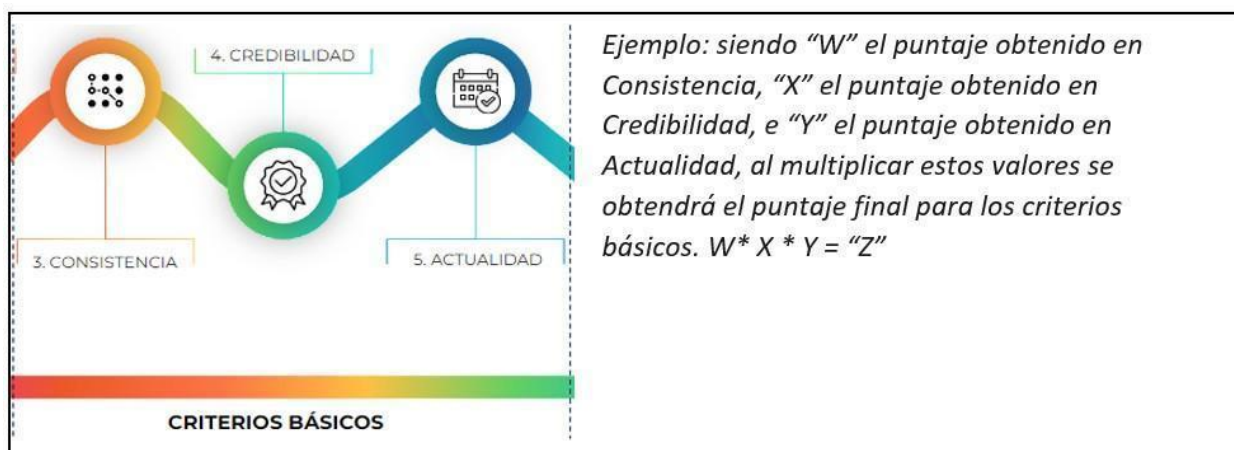
- **Paso 3:** se hará un promedio de los resultados por registro del paso 2 para obtener un puntaje total para el campo.

Resultado: para el ejemplo el porcentaje de Actualidad del campo Teléfono es del 50%.

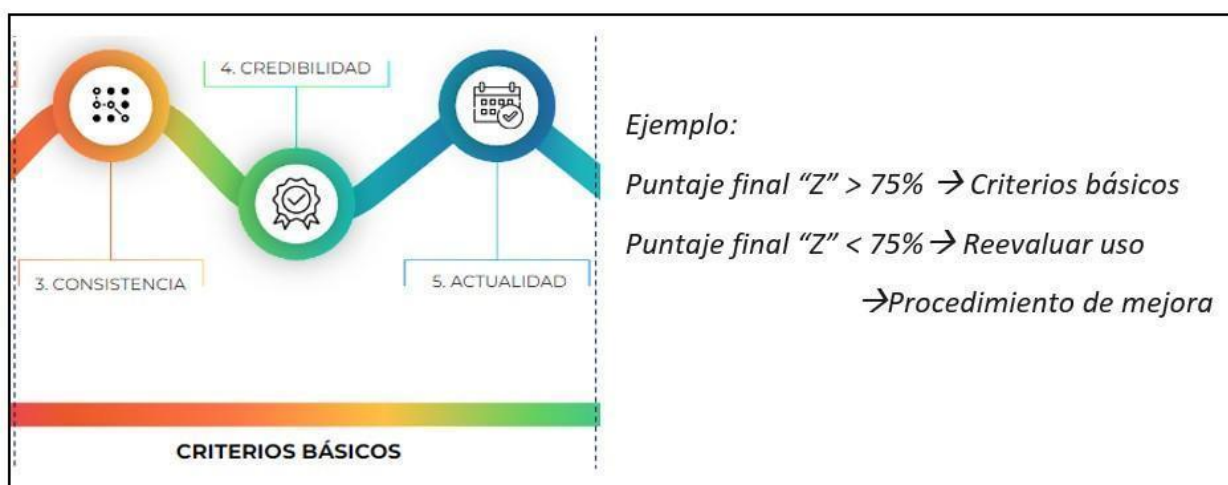
² Este campo es condición obligatoria para poder medir la actualidad.

3.2.2.4. ¿Cómo cuantificar los criterios básicos?

1. Para los criterios básicos (Consistencia, Credibilidad, Actualidad), cada conjunto de campos obtendrá un puntaje individual y el puntaje final determinará una multiplicación de los tres puntajes individuales.



2. Como parámetro general se puede indicar que un puntaje menor a 75% tanto en consistencia como en credibilidad y actualidad tomados en conjunto, indicaría que se desestime el uso del campo analizado. No obstante, si bien podemos indicar a dicho 75% como un parámetro general, será responsabilidad de cada "Dueño de Datos" definir los umbrales adecuados en función de la necesidad de cada negocio. A este parámetro lo llamaremos en adelante Umbral Límite Inferior Aceptable de Criterios Básicos (ULIAb). En el caso de incumplimiento del ULIAb sería necesario reevaluar la selección de los campos o construir procedimientos para mejorar la calidad de los datos.
3. Para aquellos campos con un puntaje total de los criterios básicos menor a 75% será necesario reevaluar su uso o construir procedimientos para mejorar la calidad de los datos.



4. Finalmente, el puntaje final de un campo estará calculado por la multiplicación de los puntajes totales de los criterios mínimos y los criterios básicos³.



3.1. Criterios óptimos

3.3.1. Pertinencia

3.3.1.1. Cualitativos

La calidad de datos implica que estos deberán coincidir con la materia o tema que se trata y que deberán contar con una concordancia para resultar útiles y que puedan ser organizados correctamente. En caso de que los datos no coincidan de manera total, es necesario contar con métodos para explicar aspectos específicos.

Es importante publicar todos los datos pertinentes, pero hay que tener cuidado de no publicar a ciegas todos los datos disponibles sin considerar su utilidad.

Por otro lado, podría preguntarse si la cantidad de datos que quiere publicar es suficiente para que los usuarios le den sentido y aporten valor, o debería añadir más datos o contexto.

3.3.1.2. Cuantitativos

La pertinencia son los datos necesarios para llevar adelante el negocio sobre los datos totales existentes.

³ Si bien el puntaje total se calcula de forma conjunta para los criterios básicos (exactitud y completitud) y mínimos (consistencia, credibilidad y actualidad), se recomienda que cada campo a nivel individual alcance el mínimo umbral de calidad de 75%.



Mal ejemplo

Nombre	Tamaño
Tráfico_2010-2015.csv	976,563 KB

El archivo en la tabla contiene tráfico ficticio de datos durante un período de 6 años. En total, el archivo tiene un tamaño de cerca de 1 GB. Si los usuarios están interesados en los datos de un año particular, igual tendrán que descargar el archivo completo



Buen ejemplo

Nombre	Tamaño
Trafico_2010.csv	97,662 KB
Trafico_2010.csv	297,833 KB
Trafico_2010.csv	228,536 KB
Trafico_2010.csv	165,139 KB
Trafico_2010.csv	39,164 KB
Trafico_2010.csv	144,886 KB

En contraste, esta tabla muestra los mismos datos separados por año. De esta forma, el tamaño de archivo se mantiene razonable y los usuarios pueden descargar los archivos exactos que necesitan. Cada archivo debería ser publicado por separado.

3.3.1.3. Pasos a seguir:

- **Paso 1:** establecer los indicadores que viabilizan el cumplimiento del caso de uso.
- **Paso 2:** cuantificar las veces que cada una de las variables individuales se involucra en el cálculo de los indicadores.
- **Paso 3:** el puntaje se calculará dividiendo la cuantificación del paso 2 sobre el total de indicadores desarrollados en el paso 1.

3.3.2. Valor agregado

3.3.2.1. Cualitativos

Son los campos complementarios, no premeditados en el indicador inicial, que profundizan la información obtenida.

El uso coherente de identificadores únicos también permite la vinculación y el aumento con datos externos. Esto añade valor a los datos existentes mediante la vinculación con nuevos conceptos o aspectos de los datos existentes.

3.3.2.2. Cuantitativos

La siguiente captura de pantalla muestra un conjunto de datos, que enumera los nombres de ingredientes cosméticos comunes con su correspondiente número de registro de resumen químico (número CAS).

Número de referencia	Nombre químico	Glosario de Nombre de ingredientes comunes	Número CAS	Número EC
1	Ácido benzoico y su sal de sodio	Ácido benzoico; benzoato de sodio	65-85-0 / 532-32-1	200-618-2 / 208-534-8
1 ^a	Sales de ácido benzoico diferentes	Benzoato de amonio	1863-63-4 / 2090-05-3 / 582-25-2	217-468-9 / 218-235-4
2	Ácido propiónico	Ácido propiónico	79-09-4 / 17496-08-1	201-176-3 / 241-503-7
3	Ácido salicílico y sus sales	Ácido salicílico/sorbato de calcio	69-72-7(1)/ 824-35-1(2)	200-712-3(1) / 212-525-4
4	Hexa -2,4- ácido dienóico y sus sales	Ácido sórbico/sorbato de calcio	110-44-1 / 7492-55-9	203-768-7 / 231-321-6
7	Bifenilo -2-ol	Fenilfenol	90-43-7	201-993-5
9	Sulfitos inorgánicos e hidrógeno	Sulfito de sodio/ bisulfito amónico	7757-83-7 / 10192-30-0	231-821-4 / 233-469-7
11	Hidrobutanol	Clorobutanol	57-15.8	200-317-6
12	Ácido hidroxibenzoico y sus sales	Ácido hidrobénzoico	99-96-7 / 9976-3	202-804-9 / 202-785-7

3.3.2.3. Pasos a seguir

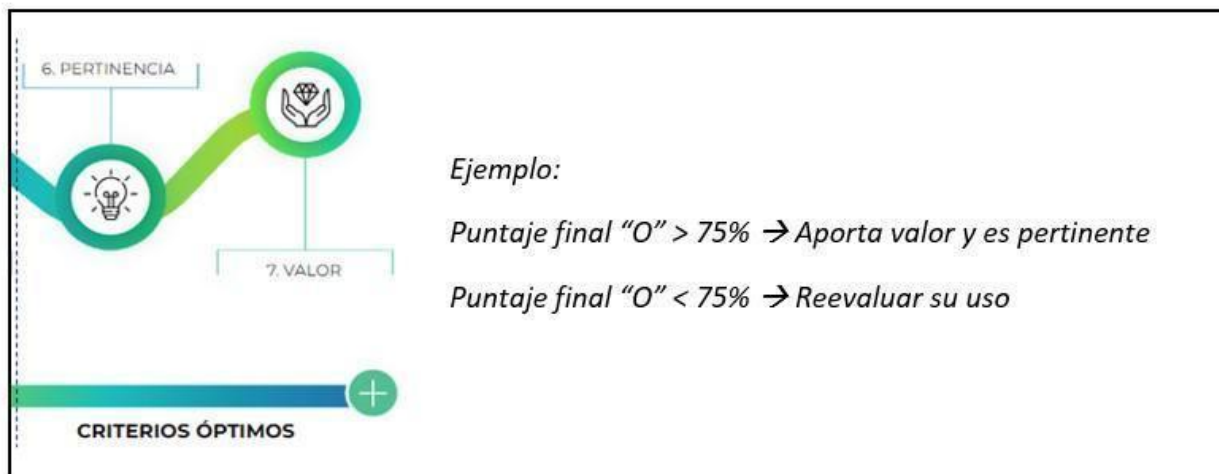
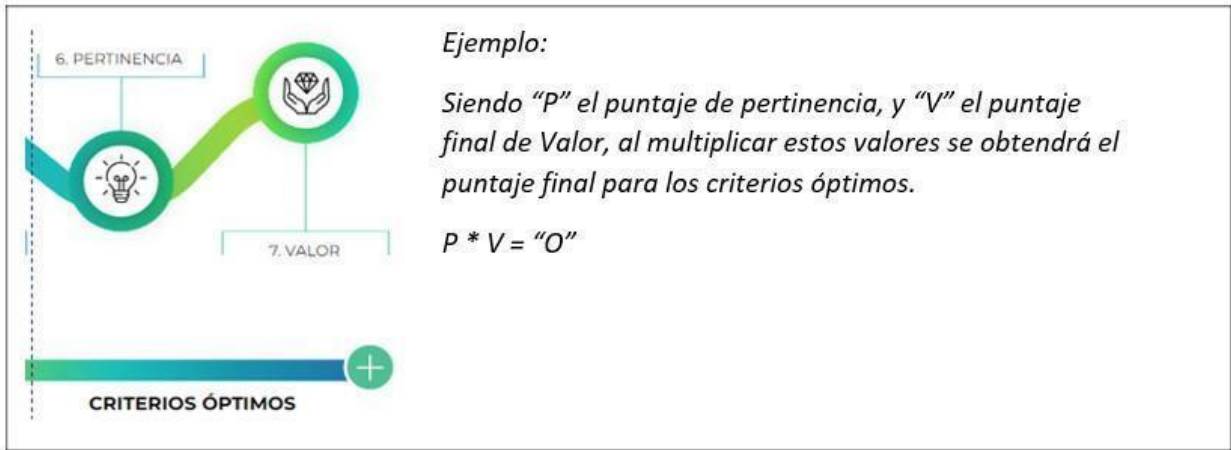
- **Paso 1:** establecer los desgloses analíticos no previstos en un inicio y que son requeridos en el desarrollo del proyecto, para profundizar y ampliar la información obtenida del indicador.
- **Paso 2:** evaluar si existen restricciones lógicas a la agregación de determinado campo a los indicadores iniciales.
- **Paso 3:** el puntaje se divide en tres:
 - El campo puede ser usados sin restricciones = 100%
 - El campo tiene restricciones de uso = 50%
 - El campo se desestima = 0% (cuando no hay registros suficientes o imposibilidades lógicas)

3.3.2.4. ¿Cómo cuantificar los criterios óptimos?

1. Para los criterios óptimos (pertinencia y Valor Agregado), cada conjunto de campos obtendrá un puntaje individual y el puntaje final lo determinará una multiplicación de los puntajes individuales.
2. Como parámetro general se puede indicar que un puntaje menor a 75% tanto en pertinencia como en valor agregado tomados en conjunto, indicaría que se desestime el uso del campo analizado. No obstante, si bien podemos indicar a dicho 75% como un parámetro general, será responsabilidad de cada “Dueño de Datos” definir los umbrales adecuados en función de la necesidad de cada negocio. A este parámetro lo llamaremos en adelante Umbral Límite Inferior Aceptable de Criterios Óptimos (ULIAo). En el caso de incumplimiento del ULIAo sería necesario reevaluar la selección de los campos o construir procedimientos para mejorar la calidad de los datos.

La cuenta se realiza:

Paso 1: por campo, evalúo pertinencia y valor agregado. Luego calculo el nivel de pertinencia y valor agregado por campo, y determinó su cumplimiento respecto de los parámetros de calidad.



4. Score de calidad

Los estándares para interpretar el porcentaje final de cada campo están medidos en el grado de confiabilidad.

Esto es relativo a cada dominio de datos y deberá ser definido por el Data Owner. (visión general que solo puede ser utilizada como marco de referencia).

Grado de confiabilidad	Tipo de confiabilidad
<u>Mayor al 90%</u>	Muy confiable
<u>Entre 75% y 90%</u>	Confiable
<u>Menor a 75%</u>	Poco confiable
<u>0%</u>	No encontrado en NINGÚN padrón confiable

El grado de confiabilidad determinará el posible uso de cada campo en la construcción de indicadores, algoritmos, consultas programadas y demás. Estos esfuerzos de construcción buscan garantizar la toma estratégica de decisiones y el alcance de objetivos operativos.