MANUAL

VERSIÓN 1.0

CALIDAD DE DATOS

GOBIERNO DE LA CIUDAD DE MÉXICO | ADIP

Elaborado por la Agencia Digital de Innovación Pública
Plaza de Las Vizcaínas 30, Centro Histórico de la Cdad. de México,
Centro, Cuauhtémoc, 06000 Cuauhtémoc, CDMX
Marzo 2021

Presentación

El presente documento fue elaborado por la Agencia Digital de Innovación Pública como parte de la implementación de la Política de Gestión de Datos de la Ciudad de México. Tiene como objetivo servir de guía a los Entes Públicos de la Administración Pública de la Ciudad de México en sus procesos de generación, recolección, análisis y publicación de datos para mejorar la calidad de los mismos.

En el presente documento se aborda el concepto de calidad de datos a lo largo del flujo de datos y las particularidades que tiene según la fuente de los datos de que se trate y se presentan recomendaciones y buenas prácticas para tener calidad de datos en cada una de las siete dimensiones de la calidad.

Contenido

Manual básico de calidad de datos

¿Qué es la calidad de datos?

La Política de Gestión de Datos de la Ciudad de México tiene por objetivo establecer las directrices generales, reglas y criterios para la gestión de los datos de los Entes de la Administración Pública, con la finalidad de facilitar y procurar su aprovechamiento en la toma de decisiones públicas y se basa en un Marco para la Gestión de Datos de la Ciudad de México (MGD-CDMX) que define y delimita su implementación.

El MGD- CDMX se conforma de ocho componentes básicos, entre los cuales se encuentra ***Calidad de datos.*** El componente de ***Calidad de datos*** se refiere a la implementación de procedimientos, prácticas y estándares para definir, controlar y mejorar la calidad de los datos, con el objetivo de reducir los riesgos y costos asociados a datos poco confiables, inconsistentes y/o imprecisos.

Los datos de baja calidad generalmente tienen poca replicabilidad y fiabilidad, lo cual dificulta su uso y en muchas ocasiones requiere largos y costosos procesos de limpieza. En numerosas ocasiones, dichos procesos de limpieza son insuficientes para garantizar el aprovechamiento de los datos.

En la Política de Gestión de Datos de la Ciudad de México, se define como calidad de datos:

> ***El grado en el cual los datos son ciertos, pertinentes, proporcionales, accesibles, íntegros y oportunos, de tal forma que satisfacen las necesidades de las personas usuarias, los tomadores de decisión, las aplicaciones y procesos que los utilizan y, en general, los objetivos institucionales para los que fueron generados y/o recolectados.***

Cada una de las dimensiones mencionadas se desagrega en factores específicos, como se muestra de forma resumida en la Tabla 1.

Tabla 1. Dimensiones y factores de la calidad de datos			
Dimensión	**Definición dimensión**	**Factores**	**Definición factor**
1. Certeza	Los datos describen el fenómeno que buscan medir de forma fidedigna, sin ambigüedad, sesgos y errores.	**1.1. Precisión de los valores de los datos.**	Los datos describen el fenómeno que buscan medir con la mayor exactitud posible.
		1.2. Certeza sobre las revisiones de los datos.	Los datos se acompañan de información metodológica y contextual explícita sobre las revisiones que puedan haber tenido.
2. Integridad	Los datos se producen de manera imparcial y se garantiza que no han sido modificados sin autorización. En caso de existir modificaciones, se proveen las aclaraciones metodológicas pertinentes.	**2.1. Imparcialidad de los datos.**	Los procesos de recolección, generación y/o análisis de datos no tiene como objetivo beneficiar o perjudicar a personas, grupos o temas particulares.
		2.2. Integridad de la fuente de los datos.	La fuente de los datos debe estar documentada y disponible para las personas usuarias
		2.3. Garantía de no modificación de los datos.	Los datos no deben ser modificados de forma arbitraria. En caso de hacerse, se debe proveer la aclaración metodológica explícita.
3. Consistencia	Los datos son coherentes entre ellos mismos y con el diseño metodológico utilizado para su generación y/o recolección, de modo que no se produce contradicción ni	**3.1. Consistencia metodológica interna.**	Los procesos de generación, recolección y análisis de datos no deben contradecirse entre ellos, así como con los objetivos de los mismos.
		3.2. Consistencia metodológica externa.	Los procesos de recolección, generación y análisis de datos deben ser pertinentes para el fenómeno que se desea medir.

	oposición entre ellos.	**3.3. Comparabilidad de los datos.**	Los datos se pueden comparar y cotejar con datos de otras fuentes similares.
4. Accesibilidad	Las personas usuarias pueden acceder, localizar, disponer, comprender y obtener los datos de forma sencilla y clara.	**4.1. Difusión de los datos por medios amplios y de fácil acceso.**	Los datos están a disposición de las personas usuarias en una variedad de medios electrónicos y físicos de fácil acceso.
		4.2. Uso de formatos y tecnologías que faciliten el uso e interpretación de los datos.	Los datos están accesibles en formatos y tecnologías de uso no propietario para un fácil acceso y uso por las personas usuarias.
		4.3. Disposición de documentos metodológicos y otros recursos de apoyo.	Los datos están acompañados por la documentación que permita su uso e interpretación por las personas usuarias.
		4.4. Contar con metadatos completos y actualizados.	Los datos cuentan con sus respectivos metadatos.
5. Oportunidad	Los datos se mantienen actualizados de tal forma que reflejan los cambios del fenómeno que buscan medir.	**5.1. Actualidad de los datos.**	El tiempo entre los cambios del fenómenos que se busca medir en el mundo real y la actualización de los datos es el mínimo posible.
		5.2. Oportunidad de entrega y/o publicación.	El tiempo entre la entrega/publicación de los datos y las necesidades de las personas usuarias es el mínimo posible.
		5.3. Puntualidad de los datos.	El tiempo entre la generación de los datos y su utilización para el fin

			para el que fueron creados es el menor posible.
6. Proporcionalidad	Los datos se producen mediantes los procesos menos costosos posibles y no son excesivos en relación con el ámbito y finalidades para los que fueron generados y/o recolectados.	**6.1. Proporcionalidad de los costos de los procesos.**	La generación, recolección y/o análisis de los datos está justificada. Los recursos económicos, materiales y humanos utilizados son suficientes pero no excesivos.
		6.2. Proporcionalidad en el rango de desagregación de los datos.	La desagregación de los datos debe ser relevante para el fenómeno que se desea medir y no invadir la privacidad de las personas sin razón.
7. Pertinencia	Los datos satisfacen las necesidades de las personas usuarias y el cumplimiento de los objetivos institucionales para los que fueron creados.	**7.1. Satisfacción de las necesidades de las personas usuarias.**	Los datos son útiles para las personas usuarias, ya sean personas del servicio público, academia o ciudadanía.
		7.2. Adecuación de los datos a los objetivos institucionales.	Los datos se alinean con los objetivos institucionales del Entes Público que los recolecta, genera o analiza y contribuyen al cumplimiento de de las políticas para los que fueron planteados.

El flujo de los datos

Los datos, desde su generación y/o recolección, hasta su uso o publicación pasan por un "flujo de datos", es decir, un proceso mediante el que se generan los datos hasta su etapa final de difusión y/o evaluación.

Las acciones que se llevan a cabo dentro de cada fase determinan la calidad de los datos. Estas fases no tienen un orden necesariamente lineal, sino que identifican las principales acciones que deben llevarse a cabo a lo largo del ciclo de vida de los datos y las relaciones entre ellas. Las acciones y las fases no son limitativas ni exhaustivas, por lo cual que se lleven a cabo depende de cada caso particular y de las necesidades de las personas usuarias.

En la tabla 2 se describe cada una de las fases del flujo de datos de manera general. Dichas fases son una versión compactada de UNECE, 2016.

Tabla 2. Flujo de los datos

	Fase	Acciones
1	**Especificación de necesidades**	Identificar una necesidad de nuevos datos o modificación de datos existentes.
		Identificar a las personas usuarias de los datos.
		Identificar las necesidades de datos de las personas usuarias.
		Definir los objetivos de los datos.
		Corroborar que dichos datos no existen ya.
		Definir el marco conceptual para su generación y/o recolección.
		Elaborar el plan general de generación y/o recolección.
2	**Diseño y construcción**	Definir, diseñar, construir y probar los instrumentos de recolección necesarios
		Definir los conceptos, metodologías y procesos operativos a utilizar.
		Definir el diseño muestral, en caso de hacerse una muestra estadística.

		Corroborar la existencia de estándares nacionales e internacionales que se puedan utilizar para reducir tiempos y costos, así como asegurar la comparabilidad e interoperabilidad de los datos.
		Definir cómo se van a procesar y analizar los datos.
3	**Generación/ recolección**	Aplicación del instrumento en la muestra o población objetivo.
		Preparación del proceso de generación o recolección.
		Ejecución de la generación o recolección.
4	**Procesamiento**	Limpiar y transformar los datos
		Integrar los datos[1]
		Clasificar y codificar los datos[2]
		Revisar los datos[3]
		Editar e imputar los datos[4]
		Derivar nuevas variables y unidades de análisis[5]
		Calcular ponderadores[6]
		Calcular agregaciones[7]

[1] Se refiere a combinar datos de una o más fuentes de datos; es decir, que provienen de distintos métodos de recolección y/o de otros conjuntos de datos, con la finalidad de obtener una versión unificada de ellos que permita utilizarlos (ver más adelante: fuente de datos).
[2] La codificación de datos es un proceso mediante el que se asignan códigos numéricos a respuestas en forma de texto por medio de un esquema de clasificación predeterminada." (Ver: UNECE, 2016 pág. 23).
[3] Se refiere al examen de los datos para identificar problemas potenciales, errores y discrepancias mediante la identificación de valores atípicos, respuestas faltantes (missing values) o errores en la codificación, entre otros.
[4] Proceso que, luego de que en la revisión de los datos se encuentren datos incorrectos, faltantes o poco confiables, se realiza para corregirlos, sustituirlos y/o eliminarlos. No existe un solo método de imputación y se pretende que se lleve a cabo cuando así lo demanden los datos. Por tanto, es imprescindible mejorar la calidad en la fase de recolección.
[5] Se refiere a la aplicación de fórmulas aritméticas a una o más variables (o atributos) para obtener nuevas variables que no fueron provistas de manera explícita en la recolección de datos.
[6] Solo aplica para datos que provienen de una fuente de datos estadísticos, como una encuesta muestral. Las ponderaciones se pueden utilizar para elevar los resultados y hacerlos representativos, o para ajustar la falta de respuesta, entre otros usos.
[7] Se refiere a el cálculo de suma de datos, medidas de tendencia central o de dispersión, entre otros.

5	**Análisis**	Preparar documentos de apoyo para las personas usuarias (notas técnicas, documentos metodológicos, diccionarios de datos, entre otros).
		Preparar borradores de resultados, como estadística descriptiva
		Validación de los datos[8]
		Interpretar y explicar resultados
6	**Difusión**	Divulgar, compartir y/o publicar los datos (dependiendo cuál sea el caso de uso)
		Dar el formato requerido a los datos
		Cargarlos en el sitio a través del cual se compartirán o descargarán.
		Documentar los metadatos
		Realizar materiales de difusión de datos (reportes, comunicados, gráficos, infografías, entre otros) y otros documentos metodológicos que se consideren necesarios para el uso y entendimiento de los datos.
7	**Evaluación**	Validar si los datos cumplen con las necesidades de los usuarios
		Validar si los datos cumplen con los objetivos planteados en la especificación de necesidades.
		Revisar si existe algún problema en alguna fase que interfiera con el cumplimiento de las necesidades y objetivos planteados.
		Modificar el proceso donde sea necesario para que los datos cumplan las necesidades y objetivos planteados.

[8] La "validación" de datos implica contrastar dichos datos de acuerdo a los objetivos definidos en las fases de Especificación de necesidades y Diseño. Las actividades de validación pueden incluir: revisar que la cobertura de la población y las tasas de respuesta son las requeridas; si aplica, verificar que los datos sean comparables; buscar y corregir alguna inconsistencia encontrada, entre otras.

Fuentes de los datos

La fuente de los datos es el método o instrumento que permite la recolección y/o generación de los mismos. Entender cuáles son las posibles fuentes de recolección y/o generación de datos es importante porque cada una de ellas contempla métodos y herramientas específicas con particularidades en términos de calidad. Pueden ser:

Primarias
Datos generados y/o recolectados directamente por la organización o Ente Público.

Secundarias
Datos que fueron generados y/o recolectados por terceros (otros Entes Públicos u organizaciones) y son reutilizados.

Las principales fuentes de datos primarias de la Administración Pública de la Ciudad de México son las que pueden observarse en la Tabla 3.

Tabla 3. Fuentes de datos primarias

	Fuente	Descripción
1	**Censos**	Son operaciones de recolección de datos de toda una población estadística en un momento determinado (INEGI). La población o universo estadístico pueden ser no sólo personas, sino también hogares, establecimientos, instituciones, eventos, fenómenos o cosas Por sus dimensiones y cobertura suelen ser muy costosos en tiempo y recursos.
2	**Encuestas muestrales**	Son operaciones de recolección de datos que captan información de una muestra estadística del universo de estudio (INEGI). Se realizan a unidades de análisis específicas según el objetivo de los datos a recabar, ya sean éstas hogares, establecimientos, instituciones, personas, eventos, fenómenos o cosas. Por sus especificidades es importante garantizar la calidad del diseño muestral.
3	**Registros Administrativos**	Esta es la principal fuente de datos de los Entes Públicos quienes producen información estadística que proviene de los datos que se integran en los trámites, servicios y acciones institucionales que

		realizan en el ejercicio de sus funciones (INEGI). Se recolectan a través de formatos de registro, bitácoras de actividades, sistemas desarrollados para realizar trámites, padrones de usuarios(as), entre otros. Aunque no existe un método estandarizado de generación y recolección de datos mediante registros administrativos, sí se pueden llevar a cabo acciones para garantizar su calidad.
4	**Mediciones automáticas a través de objetos**	Se refiere a la recolección de datos mediante instrumentos de medición físicos u objetos que generalmente están integrados con software y otras tecnologías para recolectar datos sobre ciertos fenómenos y/o actividades, entre otros. Las mediciones pueden ser sobre fenómenos físicos pero también sobre el uso que las personas le dan a los objetos. Cuando estos instrumentos de medición están conectados a otros mediante Internet de tal forma que intercambian datos, se le conoce como Internet de las Cosas (IoT por sus siglas en inglés).
5	**Datos colaterales**	Conocidos en la literatura como *exhaust data* o *data exhaust*, son los datos generados de forma colateral por las actividades, comportamiento y transacciones de las personas usuarias con productos y servicios digitales, ya sean en línea o no.

Recomendaciones y mejores prácticas de calidad de datos por dimensión

Como ya se ha mencionado, la calidad de los datos es una propiedad multidimensional, de alcance progresivo y definida en términos de varias dimensiones; es decir, de elementos interrelacionados. Cada dimensión se desagrega en *factores* específicos, de los cuáles se desprenden un conjunto de prácticas y recomendaciones deseables.

A continuación se enumeran las principales prácticas y recomendaciones de calidad de datos retomadas principalmente de UN, 2019; OCDE, 2019; AGESIC, 2020 y GDS, 2020.

Dimensión 1: Certeza	
Definición	Los datos describen el fenómeno que buscan medir de forma fidedigna, sin ambigüedad, sesgos y errores.
Preguntas rectoras	¿Se utilizan metodologías sólidas y comparables en los procesos de generación, recolección, análisis y revisión de los datos?
	¿La información sobre las metodologías utilizadas para generar, recolectar, analizar y revisar los datos está disponible para las personas usuarias?
	¿Se identifican errores, sesgos y omisiones en los datos con el objetivo de mitigarlos?
	¿Las revisiones de los datos están documentadas y explicitadas?
Factores	**Recomendaciones para tener datos certeros**
Precisión de los valores de	Identificar la presencia de valores atípicos (*outliers*), valores perdidos (*missing values*), errores de codificación (*miscoding*), entre otros para justificar y sustentar

los datos	el análisis de los datos.[9]
	Comparar datos de distintas fuentes que busquen medir el mismo fenómeno o fenómenos relacionados para verificar que no se contradigan. Si es así, identificar las causas.
	Definir de forma clara y explícita la(s) unidad(es) de medición, tanto en el conjunto de datos como en su documentación.
	En el caso de encuestas muestrales, calcular errores muestrales e identificar errores no muestrales de los datos.[10]
	En el caso de encuestas muestrales, calcular medidas de exactitud, precisión y desviación estándar de los datos.
Certeza sobre las revisiones de los datos	Cuando sea necesario publicar datos de manera preliminar, se debe hacer explícita su naturaleza no definitiva.
	Cuando los datos sufrieron revisiones y/o correcciones, hacer explícitos los periodos, las razones y la naturaleza de las mismas.

[9] Es común que se cometan errores al momento de generar o recabar datos. Por esto, es importante identificar aquellos valores que se salgan de lo común y que potencialmente indican errores, para posteriormente corregirlos. Es importante enfatizar que los valores que se salgan de lo común no son necesariamente errores, pero ayudan a identificar anomalías en los datos. Estos valores son: 1) valores atípicos (*outliers*), los cuales son observaciones que están situadas anormalmente lejos de las demás en una muestra estadística. Aunque definir si un valor es atípico o no es un ejercicio subjetivo y depende del contexto y el usuario, existen métodos estadísticos para identificarlos rápidamente; 2) valores perdidos (*missing values*), los cuales suceden cuando no se registra ningún valor para algún campo de un conjunto de datos. Cuando las causas de los valores perdidos son aleatorias y no dependen del fenómeno que se desea medir (errores en la captura, por ejemplo) se acostumbra la eliminación de los mismos y cuando la causa de los valores perdidos no son aleatorias y dependen del fenómenos que se desea medir (una persona en situación de vulnerabilidad que no quiere dar información personal, por ejemplo), es común hacer anotaciones metodológicas a los datos. 3) errores de codificación (*miscoding*), son observaciones que se clasifican o registran de forma incorrecta en el conjunto de datos.

[10] Los *errores muestrales* se presentan cuando se infieren características o atributos de una población mediante el uso de una muestra estadística (aleatoria). Dado que es imposible conocer la totalidad de una población, la medición de ciertos atributos se calculan a partir de una muestra estadística de la misma. A la diferencia entre el valor real de la población (parámetro) y el valor estimado a partir de una muestra estadística (estimador) se le denomina error muestral. Por otro lado, los *errores no muestrales* son todos aquellos que no dependen de la muestra estadística. Al ser un término sombrilla, éste incluye una enorme variedad de errores. Pueden ser aleatorios y sistemáticos. Los errores no muestrales aleatorios son impredecibles y pueden suceder por un error tanto del aparato de medición o de quienes registran los valores en un conjunto de datos. Los errores no muestrales sistemáticos son aquellos que suceden constante y sistemáticamente en un conjunto de datos. En este caso, el origen del error puede ser eliminado y corregido. Pueden ser errores de calibración en un aparato de medición o errores de un instrumento de recabación de datos, por ejemplo.

Dimensión 2: Integridad	
Definición	Los datos se producen de manera imparcial y se garantiza que no han sido modificados de forma engañosa. En caso de existir modificaciones, se proveen las aclaraciones metodológicas pertinentes.
Preguntas rectoras	¿La generación, recolección y análisis de los datos responde únicamente a los objetivos institucionales para los que fueron planeados?
	¿Los datos no han sido modificados para beneficiar o perjudicar a un grupo y/o individuos?
	¿Se indica explícitamente cualquier modificación a los datos?
	¿La fuente de los datos está claramente identificada?
Factores	**Recomendaciones para tener datos íntegros**
Imparcialidad de los datos	Garantizar que el proceso de generación, recolección, análisis y publicación de datos esté protegido de cualquier intervención externa que pueda influir en sus resultados.
	Las metodologías y técnicas de generación, recolección, análisis y publicación de datos deben basarse en criterios técnicos y/o científicos, no en juicios de valor.
	La manera de generar, recolectar, analizar o publicar datos debe evitar sesgos que puedan beneficiar o perjudicar a grupos o individuos. En caso de identificarse, deben ser mitigados.
	Poner a disposición de los usuarios los términos y condiciones al recopilar, procesar y difundir los datos, sobre todo cuando se trate de recolección de datos personales de acuerdo a lo previsto en la Ley de protección de datos personales en posesión de sujetos obligados de la Ciudad de México.
	Cuando se contraten a proveedores externos para la generación, recolección y análisis de datos, no deben existir conflictos de interés que puedan comprometer su imparcialidad por lo que se debe presentar la manifestación de no conflicto de interés por cada contrato, según lo establecido en la normatividad en la materia.
Integridad de la fuente de los datos	Evaluar y validar de manera regular la fuente de origen de los datos.
	Notificar de manera oportuna y explícita cualquier cambio a los datos de origen.
	Cuando sea necesario, el Ente Público propietario de los datos tiene derecho a comentar sobre interpretaciones erróneas y mal uso de las estadísticas por parte

	de la ciudadanía.
	Siempre que los datos provengan de otras fuentes, total o parcialmente, se deben citar los metadatos de dichas fuentes en la documentación de los datos.
Garantía de no modificación de los datos	Los datos deben estar completos y actualizados; es decir, no deben existir registros vacíos y sin información y tampoco ser un extracto de los mismos que pueda inducir a sesgos o errores.
	Establecer de manera clara las personas que con acceso a los datos, así como protocolos de seguridad que garanticen que los datos no sean modificados por personas no acreditadas,
	Corregir oportunamente y comunicar de manera explícita los errores identificados en los datos publicados.

Dimensión 3: Consistencia	
Definición	Los datos son coherentes entre ellos mismos y con el diseño metodológico utilizado para su generación y/o recolección, de modo que no se produce contradicción ni oposición entre ellos.
Preguntas rectoras	¿Los datos no se contradicen con otros datos que buscan medir el mismo fenómeno? Si es así, ¿por qué?
	¿Los datos no cambian de forma inexplicable en el tiempo?
	¿Los datos no se contradicen con las metodologías utilizadas para su generación, recolección y análisis?
	¿Los datos de distintas fuentes producen resultados similares bajo supuestos similares?
Factores	**Recomendaciones para tener datos consistentes**
Consistencia metodológica interna	Utilizar enfoques metodológicos acordes al problema de estudio y, de preferencia, retomados de otras experiencias nacionales o internacionales que estén documentadas.
	Proveer información metodológica sobre el diseño de los instrumentos de recolección y los instrumentos de recolección mismos para que puedan ser consultados por las personas usuarias.
	Cuando se lleven a cabo análisis estadísticos con los datos, examinar la validez de los supuestos del modelo (o modelos) estadístico(s) utilizado(s).
	Analizar los posibles desafíos metodológicos al utilizar datos generados por otras organizaciones, como aquellos ligados a la muestra estadística, la veracidad y/o comparabilidad de dichos datos.
	Diseñar instrumentos de recolección de datos que sean entendibles y fáciles de usar para la población de estudio para garantizar que se llegue a ella.

	Evitar hacer imputaciones de datos[11]. Cuando sean necesarias, los métodos de imputación deberán ser explícitos y estar disponibles para las personas usuarias.
	Notificar y/o anunciar con anticipación y por medios oficiales a las personas usuarias de los datos cualquier cambio en las metodologías de generación, recolección, procesamiento y/o análisis de los mismos.
	Explicitar y proveer documentación sobre cualquier cambio en las metodologías de generación, recolección, procesamiento y/o análisis de los datos.
	Estimar los efectos de cualquier cambio en las metodologías de generación, recolección, procesamiento y/o análisis de los datos.
	En el caso de encuestas muestrales, asegurarse que el diseño muestral es pertinente para explicar el fenómeno a estudiar.
	Definir una población objetivo que sea representativa del fenómeno que se quiere medir. Esta recomendación es particularmente importante para el caso de encuestas muestrales.
Consistencia metodológica externa	Utilizar conceptos y definiciones que describan de forma clara y explícita el fenómeno que se busca medir.
	Llevar a cabo pruebas de los instrumentos y metodologías de generación, recolección, procesamiento y análisis de datos para identificar posibles problemas u omisiones.
	Elaborar el flujo de los datos, en donde se indique cada una de sus fases: 1) Especificación de necesidades 2) Diseño y construcción 3) Generación/Recolección 4) Procesamiento 5) Análisis 6) Difusión 7) Evaluación

[11] La imputación estadística se refiere a la sustitución de datos no informados o no recolectados (*missing values*) por otros. Disponer de un archivo de datos completos es ideal, sin embargo, aplicar métodos de imputación inapropiados para lograrlo puede generar más problemas de los que resuelve. No existe el método de imputación ideal, sino que esta debe responder a cada situación específica. Está ampliamente documentado que la aplicación de procedimientos inapropiados de sustitución de información introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones obtenidas a partir del análisis de los datos. Para más información sobre valores perdidos (*missing values*) y métodos de imputación para datos procedentes de Encuestas en hogares puede referirse a CEPAL, 2007.

Comparabilidad de los datos	Comparar y analizar los datos con distintas periodicidades (por ejemplo, mensual, semestral y/o anual) y si existiesen diferencias, explicarlas y conciliarlas apropiadamente.
	Comparar y analizar los datos derivados de distintas fuentes y si existiesen diferencias, explicarlas y reconciliarlas apropiadamente.
	Explicar y documentar cualquier divergencia que pueda existir debido a diferencias conceptuales o metodológicas entre conjuntos de datos que buscan medir el mismo fenómeno y abarquen la misma área geográfica.
	Antes de generar o recopilar datos, analizar si es posible estandarizarlos para que sean comparables conceptual y metodológicamente con otros datos existentes, mediante el uso de catálogos de datos de referencia, definiciones, clasificaciones y unidades comunes.
	Documentar el código, modelo o proceso a través del cual se llegó a determinada integración o resultado de análisis de datos para garantizar su replicabilidad.

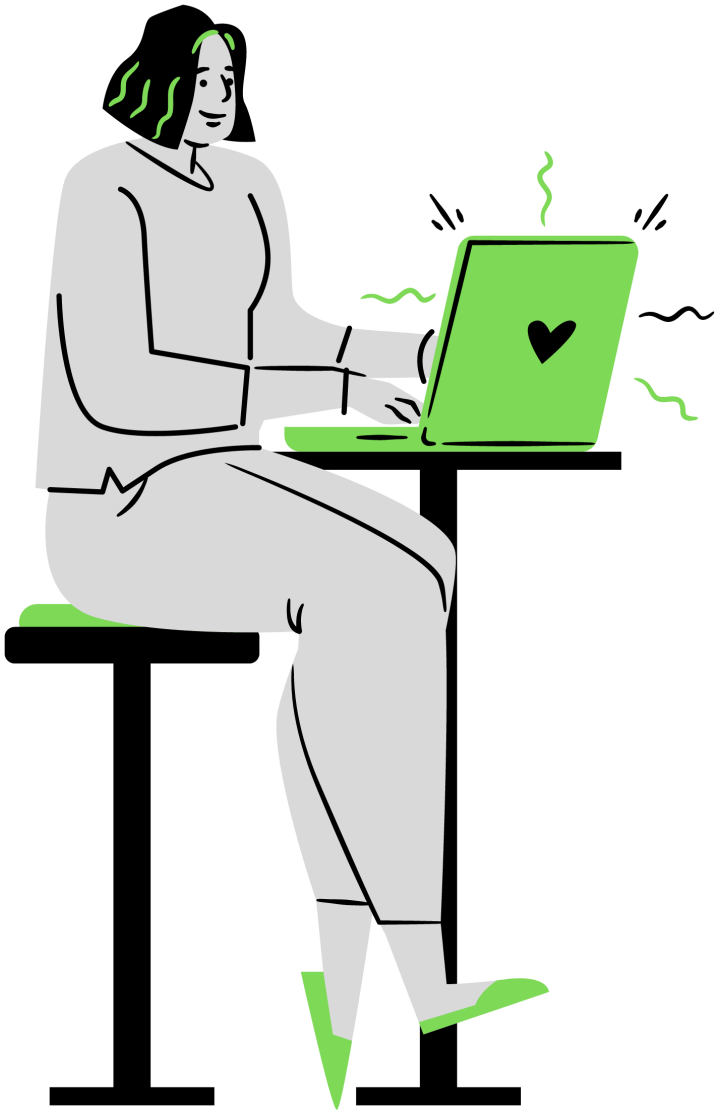
Dimensión 4: Accesibilidad	
Definición	Las personas usuarias pueden acceder, localizar, disponer, comprender y obtener los datos de forma sencilla y clara.
Preguntas rectoras	¿Los datos son legibles por máquina y por humanos?
	¿Los datos utilizan estándares abiertos siempre que es posible?
	¿Los datos se pueden encontrar rápida y fácilmente por las personas usuarias?
	¿Los datos están acompañados por documentos metodológicos para entender su contexto y significado?
Factor	**Recomendaciones para tener datos accesibles**
Difusión de los datos por medios amplios y de fácil acceso	Los datos deben estar disponibles para las personas usuarias mediante un rango amplio de medios, principalmente la descarga por Internet y APIs/servicios web.
	El sitio web del Ente Público y el Portal de Datos Abiertos de la Ciudad de México deben ser los puntos centrales de disposición y publicación de los datos públicos que se abran a la ciudadanía.
	Los datos deben publicarse de manera gratuita, con la excepción de aquellos que deban producirse a pedido, de acuerdo a lo establecido en la Ley de Transparencia, Acceso a la Información Pública y Rendición de Cuentas de la Ciudad de México y demás normatividad en la materia.
Uso de formatos y tecnologías que faciliten el uso e interpretación de los datos	Los datos deben compartirse y publicarse en formatos abiertos y legibles por máquina, tales como CSV, TSV, JSON, SHP, XML, HTML, SQL, entre otros.
	Los datos deben publicarse con una licencia CC0 (Creative Commons) de tal forma que puedan ser usados libremente por las personas usuarias con la única condición de citar la fuente.
	Priorizar el uso de Interfaces de Programación de Aplicaciones (APIs) y Servicios Web para compartir los datos.
	Utilizar herramientas que facilitan que personas con cualquier tipo de discapacidad puedan acceder a los datos, con base en las *Medidas básicas de accesibilidad y ajustes razonables en la información y las comunicaciones*, publicadas por el Instituto de las Personas con Discapacidad de la Ciudad de México (INDISCAPACIDAD).

	Generar, recolectar y almacenar los datos con la mayor granularidad posible, tomando en cuenta la dimensión de proporcionalidad.
	Siempre que sea aplicable, desagregar los datos por sexo, género, autoadscripción y condición de habla indígena, discapacidad, y todas aquellas desagregaciones que promuevan la visibilización de grupos prioritarios para el gobierno de la Ciudad de México.
	Consultar a la ciudadanía de manera periódica para conocer los formatos de difusión que más prefieren.
Disposición de documentos metodológicos y otros recursos de apoyo	Elaborar guías o manuales para los publicadores de datos, donde se definan los criterios mínimos de publicación, las estructuras, formatos y estilos preferidos con el fin de estandarizarlos.[12]
	Los datos deben estar acompañados de documentos metodológicos actualizados que ayuden a la interpretación y análisis.
	Los datos deben estar acompañados por su respectivo diccionario de datos que permita el entendimiento de cada uno de los campos o variables.
	Los datos deben estar acompañados de un servicio de soporte rápido mediante al menos correo electrónico para resolver aclaraciones o consultas de los usuarios.
Contar con metadatos completos y actualizados	Todos los datos, independientemente de su publicación o compartición, deben estar acompañados por sus respectivos metadatos completos y actualizados.
	Contar con un catálogo de metadatos sistematizado que permita la búsqueda, acceso y utilización de los datos.
	Usar conceptos, unidades, variables, catálogos de referencia y otras clasificaciones de uso común y compartido para que sean estandarizados y conocidos por todas las personas al interior del Ente Público.

[12] Por ejemplo: INEGI (2010). Presentación de datos estadísticos en cuadros y gráficas.

Dimensión 5: Oportunidad	
Definición	Los datos se mantienen actualizados de tal forma que reflejan los cambios del fenómeno que buscan medir.
Preguntas rectoras	¿La periodicidad de los datos refleja los cambios del fenómeno que buscan medir?
	¿Los datos están disponibles cuando se necesitan para tomar decisiones?
	¿Se cumplen las fechas de entrega o publicación de los datos?
Factores	**Recomendaciones para tener datos oportunos**
Actualidad de los datos	Indicar de forma explícita la última actualización de los datos.
	Los datos deben tener una frecuencia de actualización y/o publicación programada explícita disponible para las personas usuarias.
	Los datos deben cumplir con la frecuencia de actualización y/o publicación programada.
	La desagregación temporal de los datos debe reflejar los cambios del fenómeno que se busca medir.
	La desagregación temporal de los datos debe ser explícita en los nombres de los campos y/o variables (si los datos están desagregados de forma diaria, semanal, quincenal, mensual, trimestral, anual, sexenal, etc.)
	Indicar de forma explícita la cobertura temporal de los datos, es decir, el periodo de tiempo que cubren.
	Determinar una calendarización explícita de revisiones a los datos para posteriormente hacer las correcciones pertinentes.
Oportunidad de entrega y/o publicación	Consultar a las personas usuarias de los datos para la determinación de la periodicidad de actualización y desagregación temporal de los datos.
	La entrega y/o publicación de los datos debe ser lo más inmediato posible después de su generación y/o procesamiento para garantizar que los datos sean vigentes.
	Cumplir con las fechas de publicación de los datos de acuerdo al calendario de difusión. En caso de retraso, se deberán comunicar los motivos.
	Notificar por adelantado cualquier cambio en el calendario previsto de difusión, en el que se aclaren los motivos y se establezca una nueva fecha.

	Cuando se publiquen datos preliminares, verificar que cumplan con características mínimas de calidad para que sean de utilidad, señalando de manera explícita su naturaleza preliminar.
Puntualidad de los datos	Registrar el tiempo utilizado en los procesos de generación, recolección, procesamiento y análisis de los datos.
	Las fechas de entrega o publicación de los datos son suficientes para no comprometer la calidad de los mismos.



Dimensión 6: Proporcionalidad	
Definición	Los datos se producen mediante los procesos menos costosos posibles y no son excesivos en relación con el ámbito y finalidades para los que fueron generados y/o recolectados.
Preguntas rectoras	¿Los fines de la generación, recolección y/o análisis de los datos están debidamente justificados?
	¿Los costos de generar, recolectar y/o analizar los datos son eficientes en tiempo y recursos?
Factores	**Recomendaciones para tener datos proporcionales**
Proporcionalidad de los costos de los procesos	Documentar y revisar periódicamente los costos de generación, recolección y análisis de los datos con el objetivo de evaluar si son los mejores y más eficientes.
	Justificar cada campo y/o variable en los instrumentos de generación y/o recopilación de datos para evitar que éstos sean excesivos en términos de costos y privacidad de las personas.
	Diseñar los instrumentos de generación y recopilación de datos de manera tal que permitan minimizar el costo y el tiempo de codificación, preparación y limpieza de los datos.
	Utilizar diferentes medios de generación y recolección de datos que reduzcan los costos sin comprometer la calidad de los datos, como los formularios digitales.
	Antes de iniciar un nuevo proyecto de generación y recolección de datos, evaluar el uso de fuentes existentes de datos, como censos y encuestas, u otros datos provenientes de otros Entes Públicos, dependencias federales, Instituciones educativas entre otros.
Proporcionalidad en el rango de desagregación de los datos	Al publicar datos hacerlo con la mayor desagregación posible sin vulnerar la privacidad de terceros o poner en riesgo las actividades del Gobierno de la Ciudad de México.
	En caso de que el uso de los datos pudiera interferir con el derecho a la privacidad de las personas, analizar otras maneras menos invasivas de lograr los objetivos.
	Informar a los usuarios u organizaciones que proveen los datos el aviso de privacidad, los términos y condiciones de uso y obtener el consentimiento informado de los mismos.

Dimensión 7: Pertinencia	
Definición	Los datos satisfacen las necesidades de las personas usuarias y el cumplimiento de los objetivos institucionales para los que fueron creados.
Preguntas rectoras	¿Para qué se generan, recolectan y/o analizan los datos?
	¿Quiénes son las personas usuarias de los datos?
	¿Los datos son de utilidad para las personas usuarias?
	¿Cómo contribuyen los datos al cumplimiento de objetivos institucionales?
Factores	**Recomendaciones para tener datos pertinentes**
Satisfacción de las necesidades de las personas usuarias	Implementar procesos de consulta periódica para conocer e identificar nuevas necesidades de datos de las personas usuarias y darles a conocer los resultados.
	Medir y dar seguimiento de manera regular al grado de satisfacción y cobertura de las necesidades de las personas usuarias de los datos.
	Describir las limitaciones que pudieran tener los datos, para que sean tomadas en cuenta para futuras políticas o servicios que utilicen estos datos como evidencia.
Adecuación de los datos a los objetivos institucionales	Justificar debidamente el uso y objetivo de los datos, realizando análisis sobre la adecuación y necesidad de los mismos para lograr los objetivos propuestos.
	Realizar revisiones a los principales instrumentos metodológicos que producen datos de manera periódica para evaluar si se adhieren a los objetivos institucionales y, en su caso, a estándares internacionales.
	La recopilación de cualquier elemento o variable que sea idéntica o similar a las recopiladas en otros instrumentos se deberá limitar a lo que se considere necesario para la verificación o integración con los datos pre-existentes.

Referencias

Agencia de Gobierno Electrónico, Sociedad de la Información y Conocimiento (AGESIC) (2020). Marco de referencia para la gestión de calidad de datos. Gobierno de Uruguay. Recuperado de: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/marco-referencia-para-gestion-calidad-datos

Comisión Económica de las Naciones Unidas para Europa (UNECE) (2016). Modelo genérico del proceso estadístico (GSBPM), Versión 1.0 en español, septiembre de 2016. Versión en español recuperada de: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.58/2016/mtg4/Paper_8_GSBPM_5.0_v1.1.pdf

Comisión Económica para América Latina (CEPAL) (2007). Imputación de datos: teoría y práctica. Documento preparado por Fermando Medina y Marco Galván. CEPAL- División de Estadística y Proyecciones económicas. Recuperado de: https://repositorio.cepal.org/bitstream/handle/11362/4755/S0700590_es.pdf

Gaceta Oficial de la Ciudad de México (3 de diciembre de 2020). *Aviso por el que se hace del conocimiento del público en general y a la Administración Pública de la Ciudad de México, el enlace electrónico, en el que puede consultarse las medidas básicas de accesibilidad y ajustes razonables en la información y las comunicaciones.* Vigésima Primera Época, No. 486, pág. 42. Recuperado de: https://data.consejeria.cdmx.gob.mx/portal_old/uploads/gacetas/98db239c1e634244c49aacede6a807b0.pdf

Government Digital Service (GDS) (2020). Data Ethics Framework. UK Government. Recuperado de: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/917805/Data_Ethics_Framework.pdf

Instituto Nacional de Estadística y Geografía (2010). Presentación de datos estadísticos en cuadros y gráficas. Serie: Documentos técnicos para la generación de estadística básica. México, INEGI. Recuperado de: http://internet.contenidos.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/metodologias/dtgeb/Presen_cuadros_graficas/Presen_cuadros_graficas.pdf

Instituto Nacional de Estadística y Geografía (s/f). *¿Quiénes somos? Generador de información estadística*. Recuperado el 19 de marzo de 2021 de: https://www.inegi.org.mx/inegi/contenido/infoest.html

Organización para la Cooperación y el Desarrollo Económico (OCDE) (2019). Recomendaciones del Consejo de la OCDE sobre buenas prácticas estadísticas, Instrumentos Jurídicos de la OCDE. Versión en español recuperada de: https://www.oecd.org/statistics/good-practice-toolkit/OECD-LEGAL-0417-spa.pdf

Statistics Finland (2007). Quality Guidelines for Official Statistics, 2nd Revised Edition, Multiprint, Helsinki.

United Nations (UN) (2019). United Nations Nacional Quality Assurance Framework Manual of Official Statistics, Department of Economic of Social Affairs. Recuperado de: https://unstats.un.org/unsd/methodology/dataquality/un-nqaf-manual/

GOBIERNO DE LA CIUDAD DE MÉXICO | ADIP

Elaborado por la Agencia Digital de Innovación Pública
Plaza de Las Vizcaínas 30, Centro Histórico de la Cdad. de México,
Centro, Cuauhtémoc, 06000 Cuauhtémoc, CDMX
Marzo 2021