



**El futuro  
es de todos**

**DNP**  
Departamento  
Nacional de Planeación



# LEILA – Librería de calidad de datos

Unidad de Científicos de Datos  
 Dirección de Desarrollo Digital

Agosto, 2020



# Agenda

1. Introducción
2. El Portal de Datos Abiertos
3. La calidad de los datos
4. LEILA
5. Repositorio de GitHub

# 1. Introducción

# La Unidad de Científicos de Datos

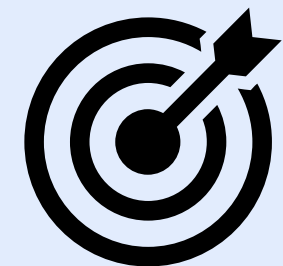
¿Quiénes  
somos?

Equipo de la Dirección de Desarrollo Digital del DNP dedicado a la explotación de datos para la formulación de políticas públicas en Colombia



¿Qué  
hacemos?

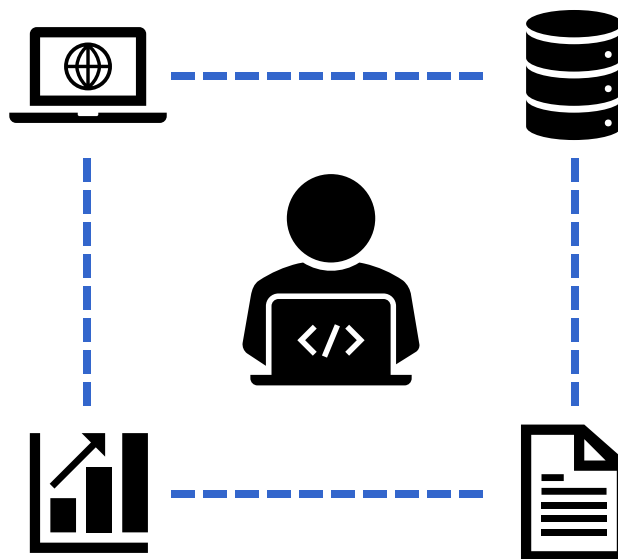
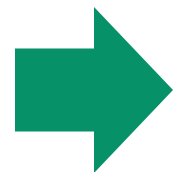
Utilizamos la analítica de datos para extraer valor de los datos y brindar insumos que orienten la toma de decisiones



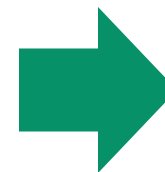
# ¿Cómo trabaja la UCD?



Datos estructurados y no estructurados



Unidad de Científicos de Datos



Apoyo a la toma de decisiones del sector público

- Objetivas
- Basadas en datos

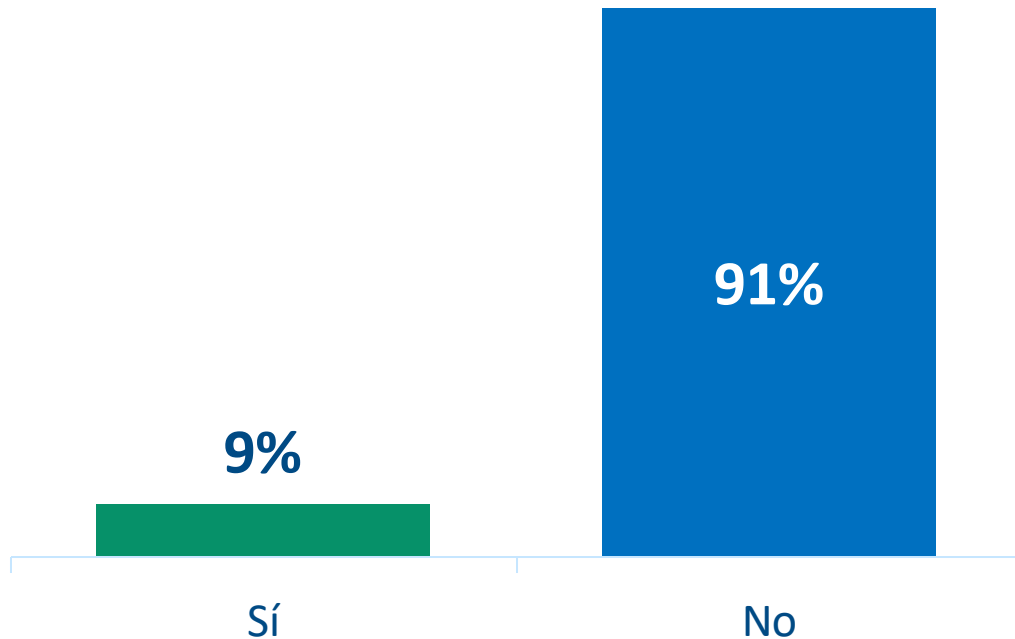
# El CONPES de Big Data\*

## IV. Marco jurídico, ético e institucional

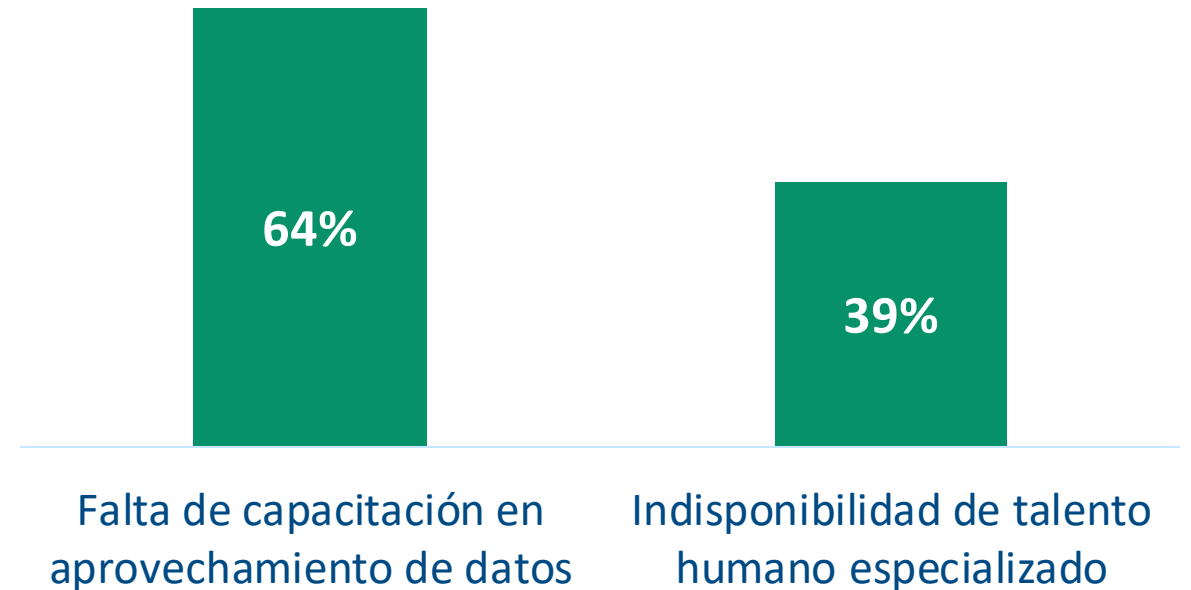


# Capacidades: bajo capital humano en 2017

Entidades con al menos un proyecto de explotación de datos



Barreras para la implementación de explotación de datos en las entidades (respuesta múltiple)






Fuente: Encuesta DNP (2017). "Sí" incluye únicamente a las entidades respecto de las que se validó el uso de algoritmos.



# Motivación: Librería de Calidad de Datos

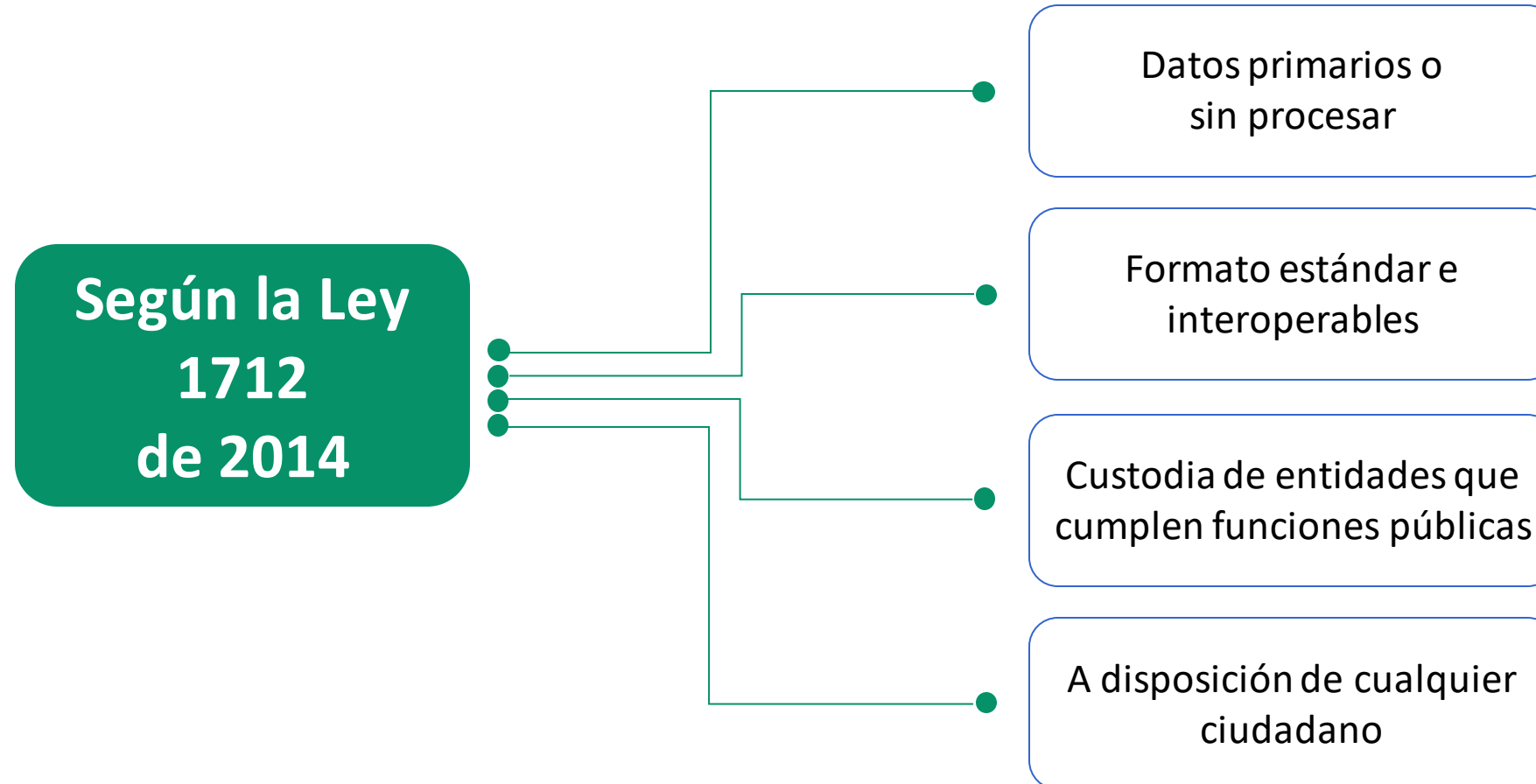


# LEILA: Librería de calidad de datos

<p>¿Qué?</p>	<p>Se desarrolló una librería para calcular métricas descriptivas y de calidad de datos para conjuntos de datos estructurados</p>	
<p>¿Para qué?</p>	<p>Para facilitar el uso de datos estructurados, incluyendo Datos Abiertos, y mejorar capacidades de análisis de datos en ciudadanos y entidades públicas</p>	
<p>Avance</p>	<p>Primera versión publicada en GitHub</p>	

# 2. El Portal de Datos Abiertos

# ¿Qué son los datos abiertos?



# El Portal de Datos Abiertos de Colombia

¿Qué es?	Portal web del Gobierno de Colombia
¿Qué contiene?	Datos abiertos
¿Quién publica?	Entidades con funciones públicas
¿Desde cuándo?	2016



[datos.gov.co](https://datos.gov.co)



# Las bases de datos del Portal

## Tipos de contenido

- Conjuntos de datos
- Enlaces externos
- Geográficos
- Gráficos
- Otros

## Sectores

- Agricultura
- Comercio
- Turismo
- Transporte
- Ciencia
- Cultura
- Otros

## Entidades

- Alcaldías
- Agencias
- Superintendencias
- Universidades

Alrededor de **20.000** publicaciones

# Uso de Datos Abiertos



¿Qué tanto se usa  
realmente el portal?

¿Qué tan buena es la calidad de los datos allí  
publicados?

# Descargas del Portal

21%

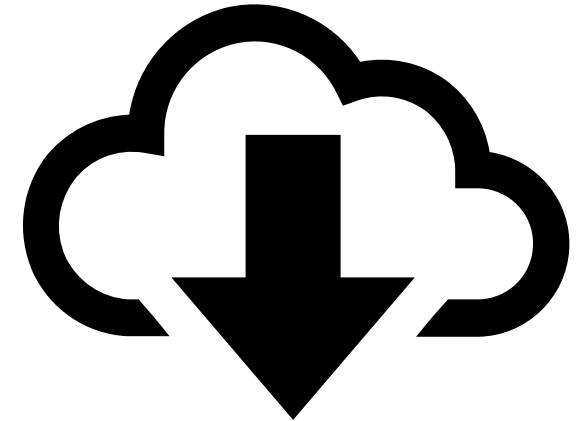
del contenido no ha sido descargado

7%

de los conjuntos de datos no han sido descargados

38%

de los conjuntos de datos no han sido descargados más de 20 veces





# Conjuntos de datos más descargados

Conjunto de datos	Número de descargas*
Casos positivos de COVID-19 en Colombia	423.000
Rutas de transporte urbano	52.000
Código único de medicamentos vigentes	31.000

Números aproximados, obtenidos el 12 de agosto de 2020

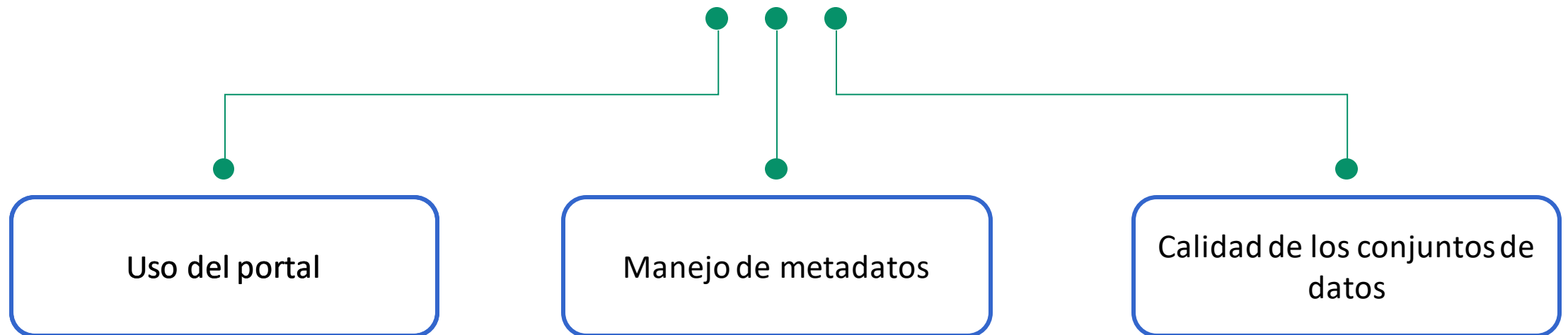
Fuente: tabla de inventario de datos.gov.co

# 3. La calidad de los conjuntos de datos

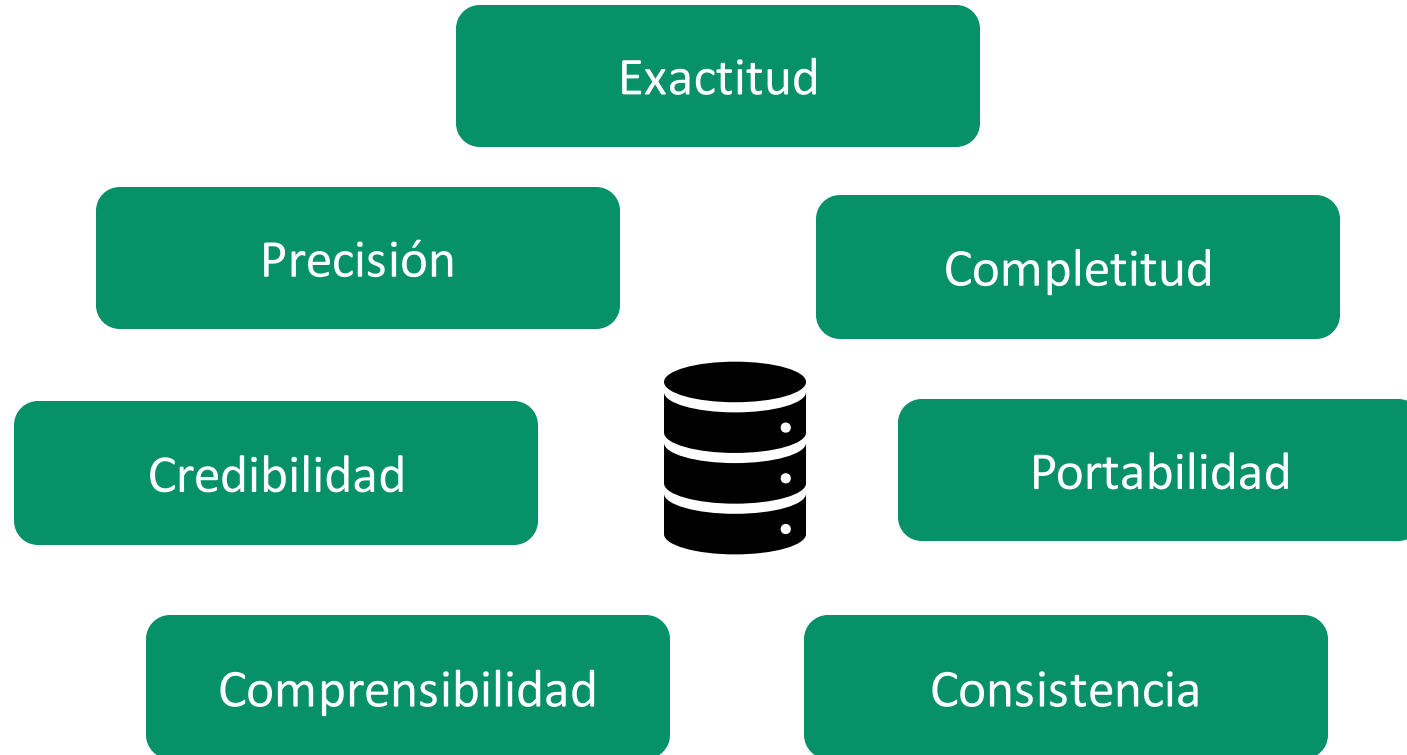
# Crerios de calidad e interoperabilidad

“Guía de estándares de calidad e interoperabilidad de los datos abiertos del Gobierno de Colombia” de MinTIC

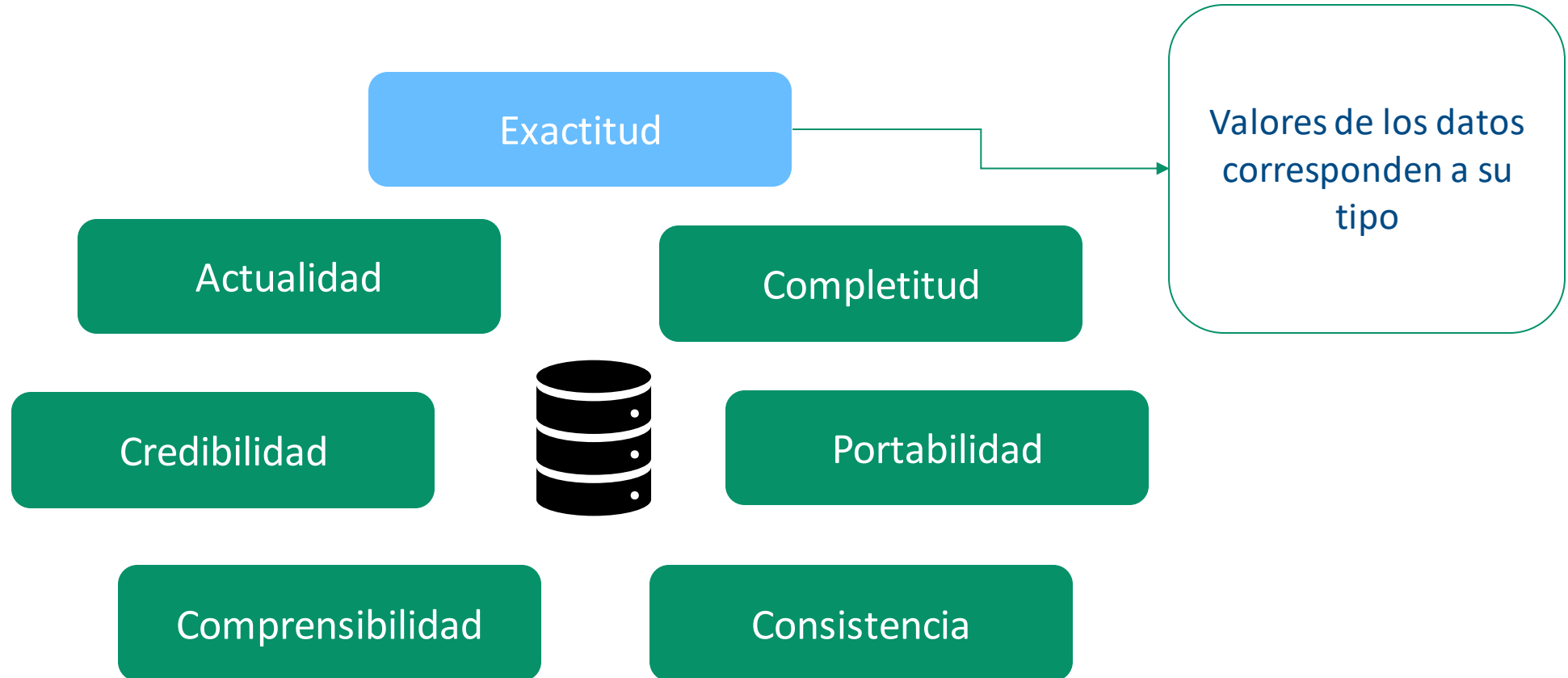
## 16 estándares de calidad e interoperabilidad



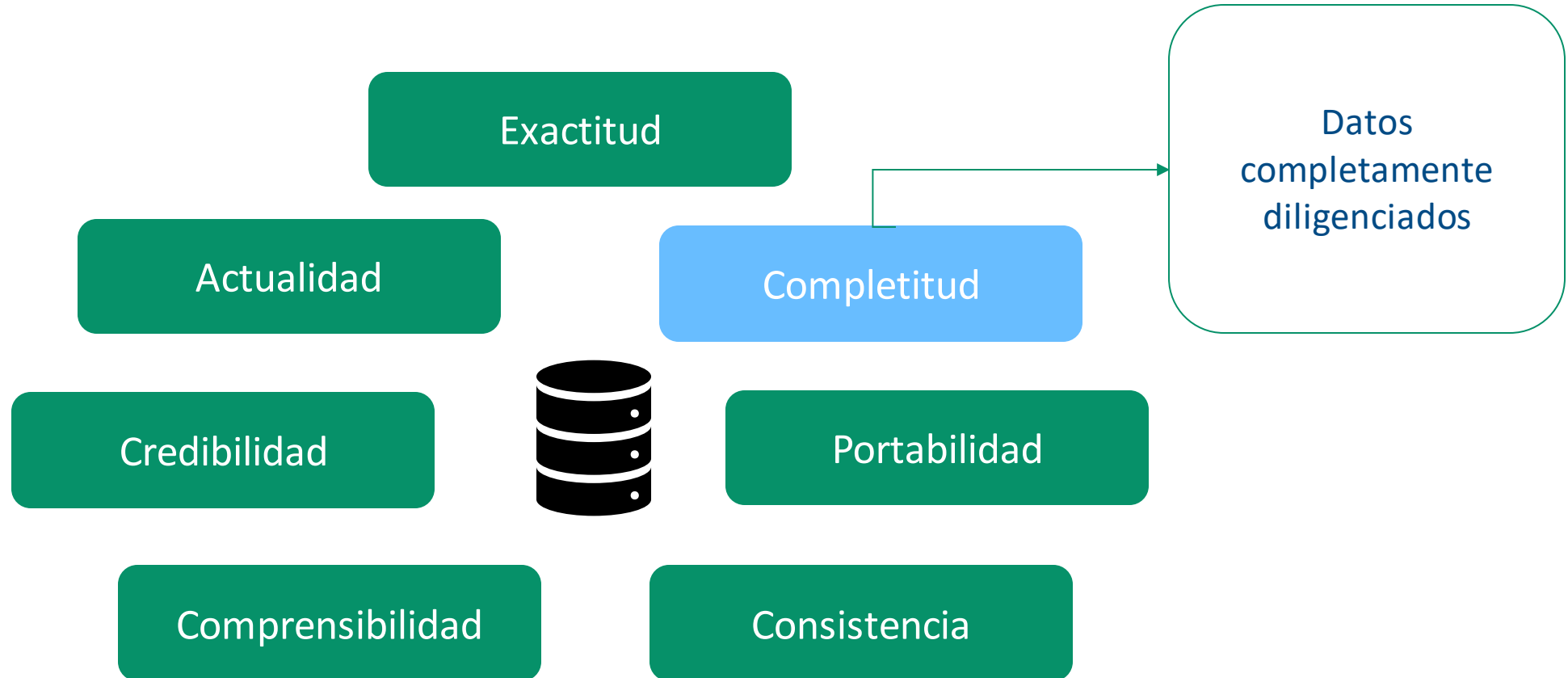
# Calidad de los conjuntos de datos



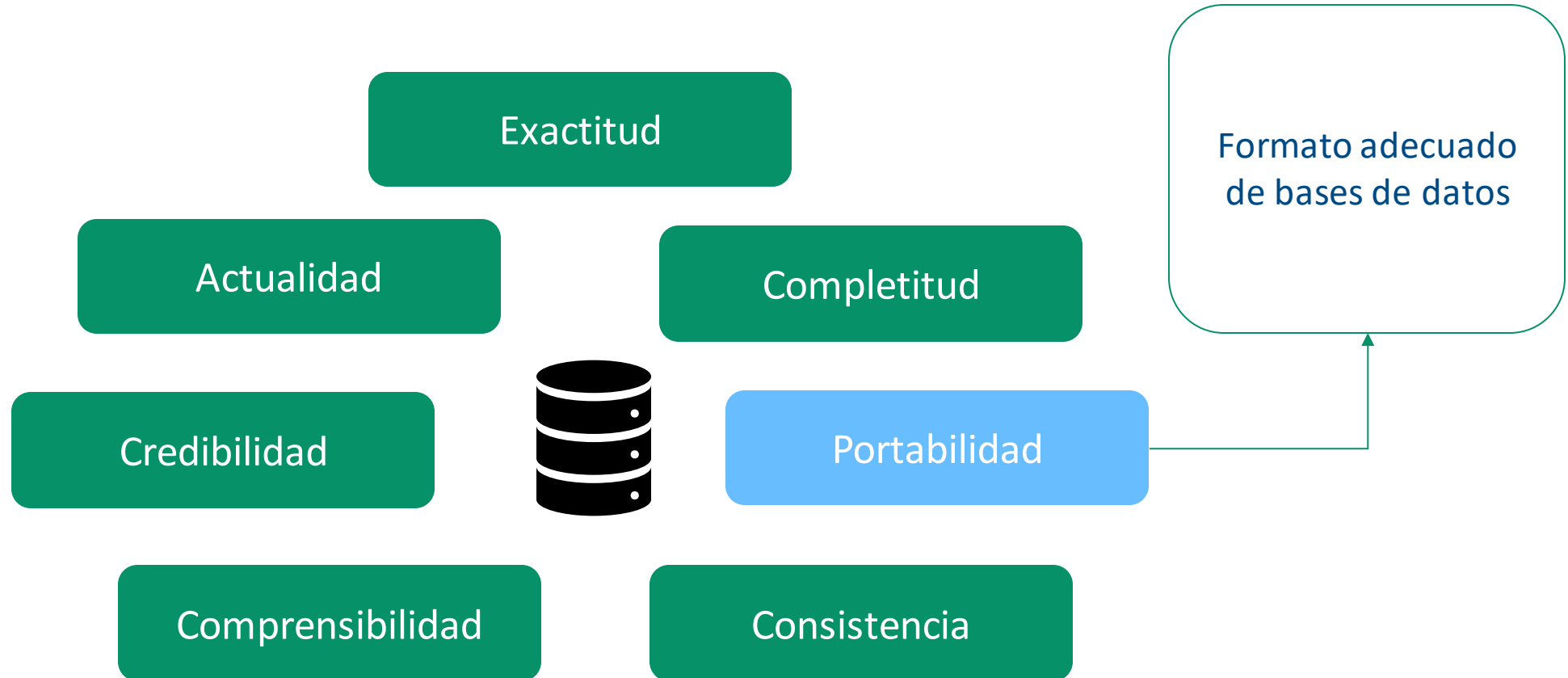
# Calidad de los conjuntos de datos



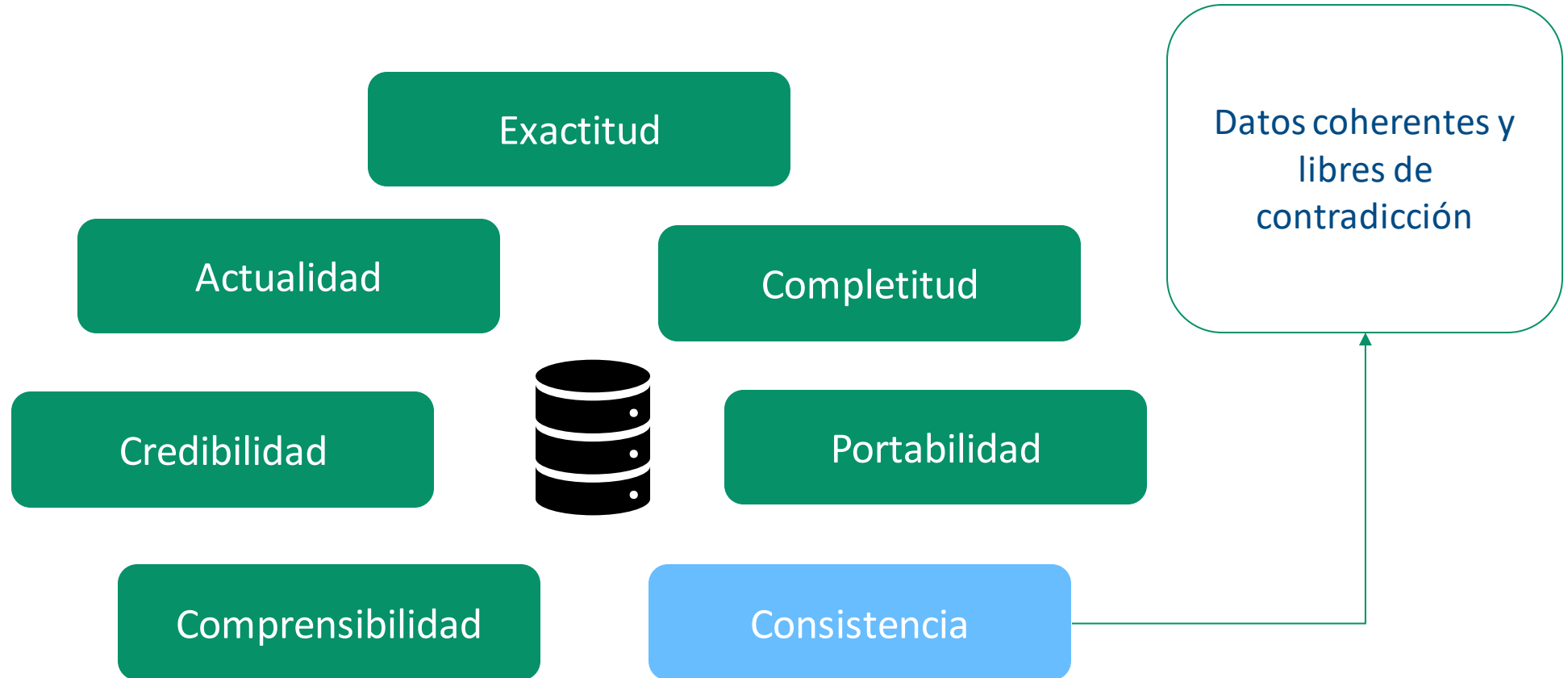
# Calidad de los conjuntos de datos



# Calidad de los conjuntos de datos

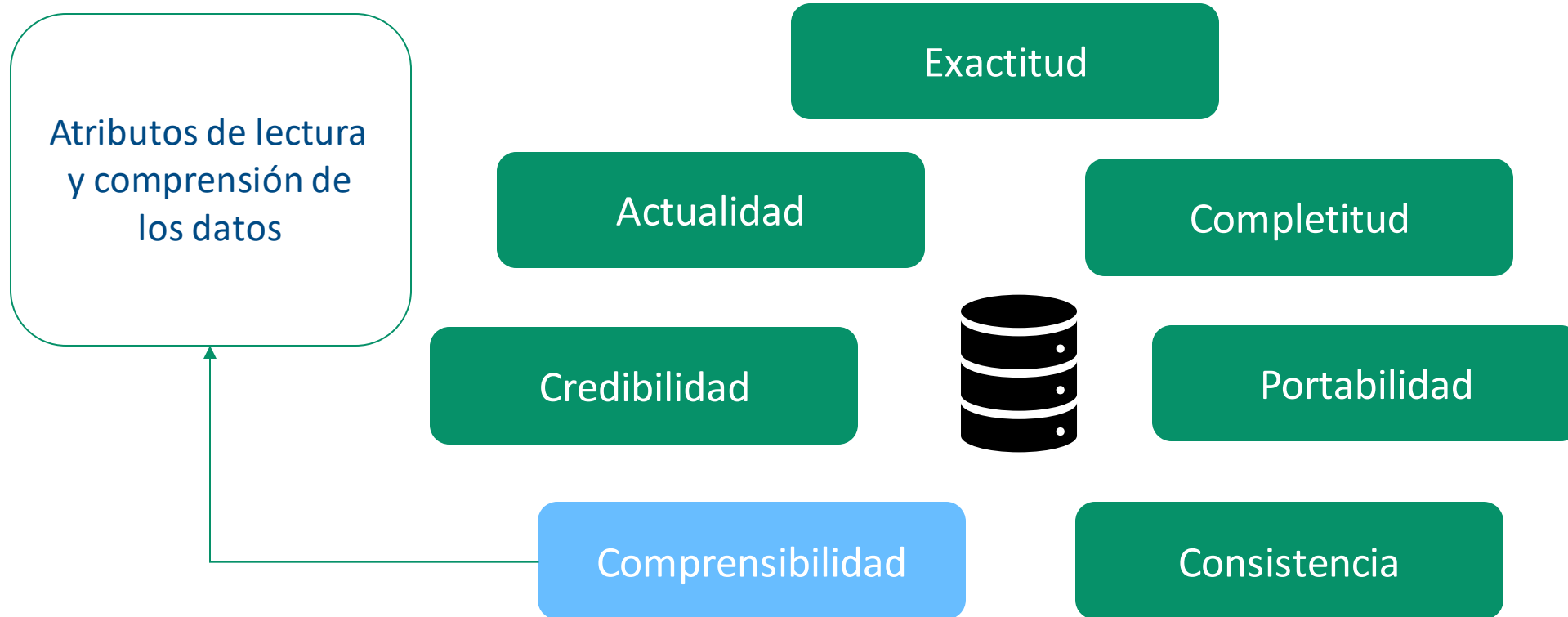


# Calidad de los conjuntos de datos

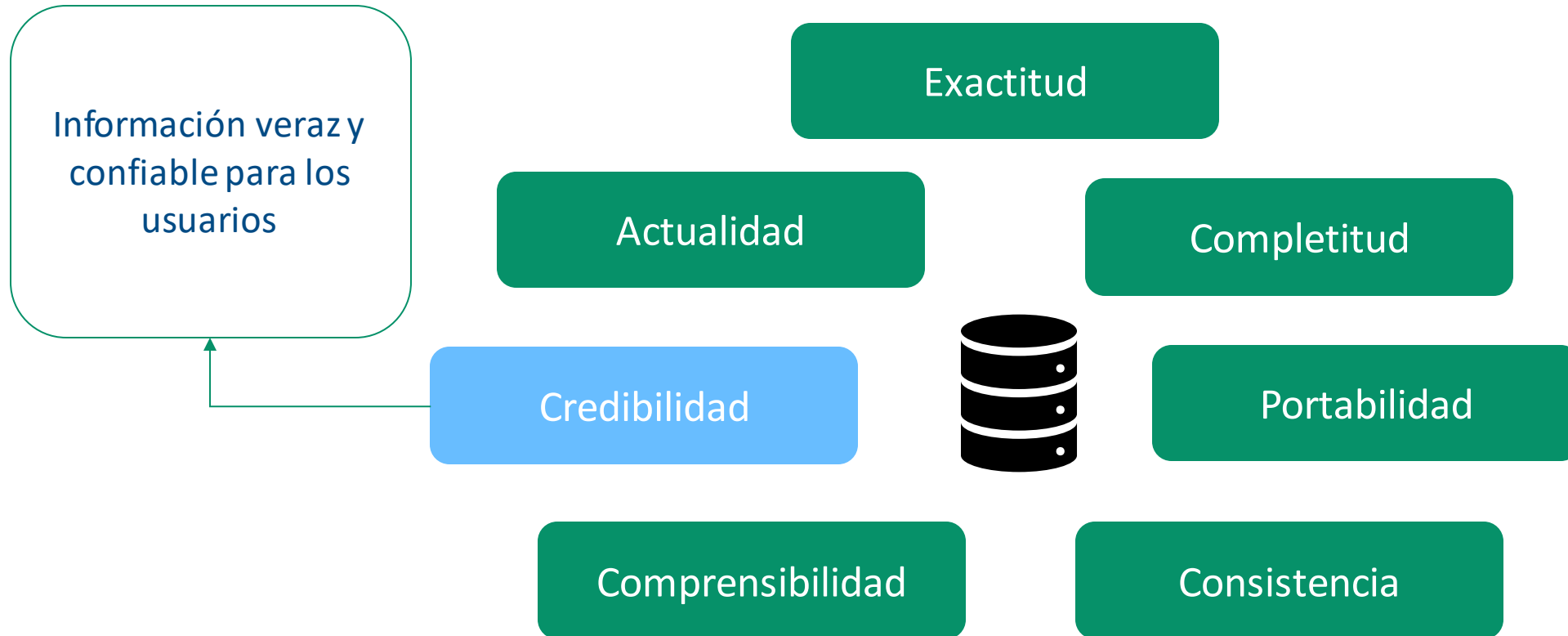




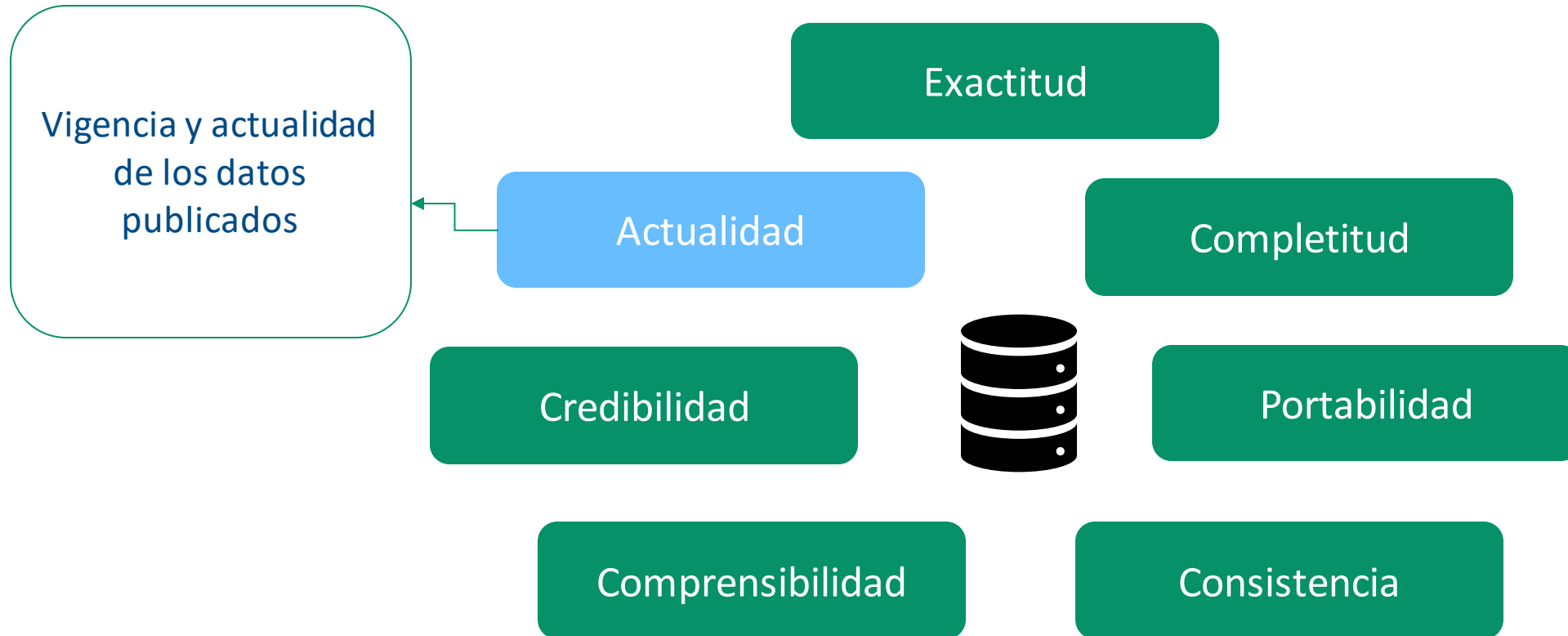
# Calidad de los conjuntos de datos



# Calidad de los conjuntos de datos



# Calidad de los conjuntos de datos



# ¿La calidad es objetiva?



Los análisis “objetivos” de calidad de datos solo pueden llegar hasta un punto

# 4. LEILA – Librería de análisis de calidad de datos

# ¿Qué es la librería LEILA?



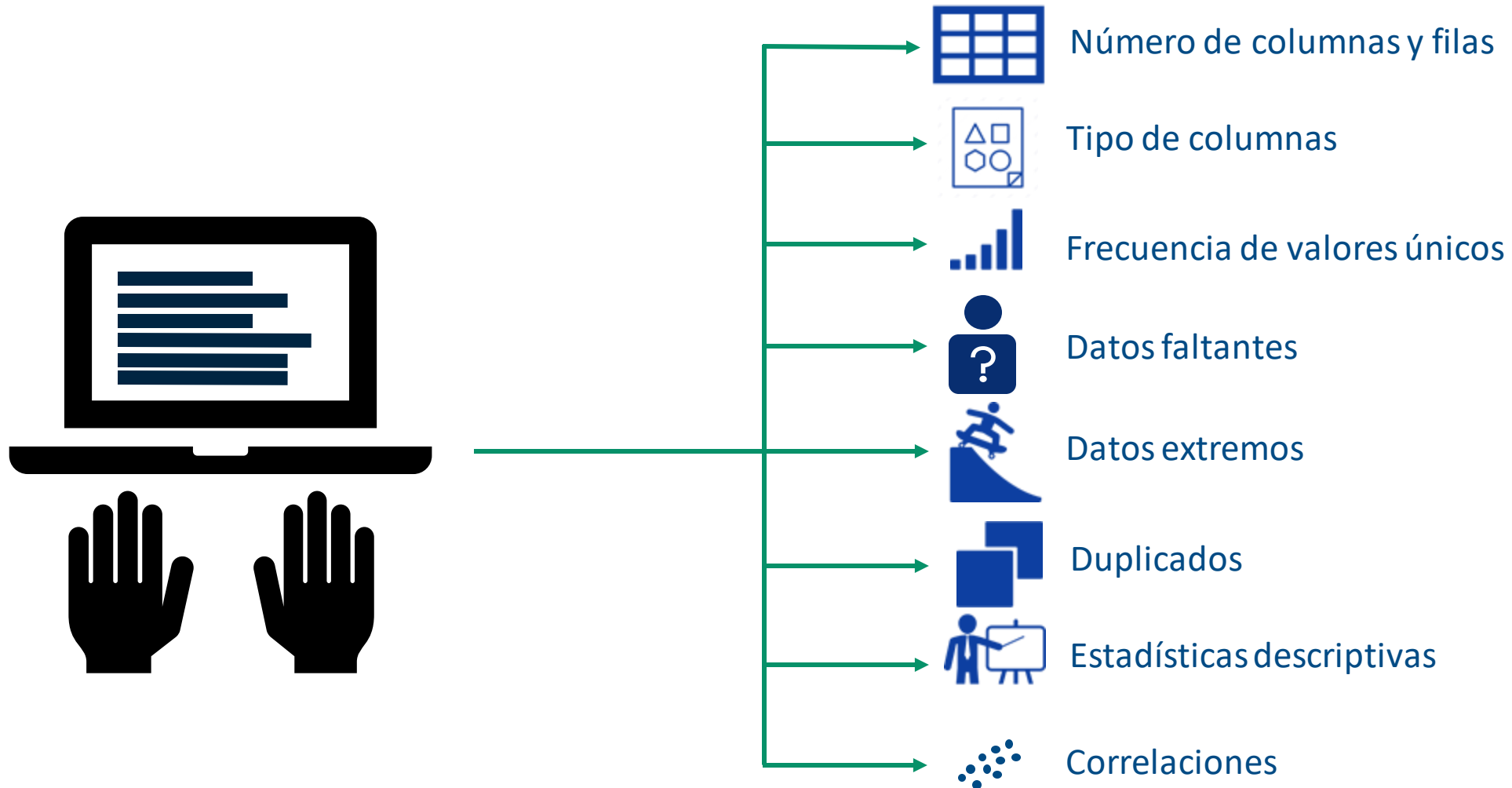
- Es una librería abierta desarrollada en Python
- Facilita la verificación de contenido y el cálculo de métricas de calidad de datos
- Funciona para cualquier conjunto de datos estructurado

# Módulos de la librería

- 1** **calidad\_datos**: calidad de los datos
- 2** **datos\_gov**: conexión a [datos.gov.co](https://datos.gov.co)
- 3** **reporte**: reporte automático

# Módulo calidad\_datos

Calcula métricas de descriptivas y de calidad para cualquier conjunto de datos.





# Módulo calidad\_datos

## *Estadísticas generales*

Categoría	Valor
Número de filas	65,616
Número de columnas	24
Columnas numéricas	7
Columnas de texto	11
Columnas booleanas	0
Columnas de fecha	6

Categoría	Valor
Otro tipo de columnas	0
Número de filas repetidas	10
Columnas con más de la mitad de datos faltantes	4
Columnas con más del 10% de datos como extremos	0
Uso en memoria de la base en megabytes (aproximado)	12

# Módulo calidad\_datos

## Tipo de cada columna

Variable	Tipo general	Tipo general (Python)
ID de caso	Numérico	int64
Fecha de notificación	Fecha	datetime64[ns]
Código DIVIPOLA	Numérico	int64
Ciudad de ubicación	Texto	object
Departamento o Distrito	Texto	object
Atención	Texto	object
Edad	Numérico	int64
Edad 2	Numérico	int64
Edad meses	Numérico	int64
Sexo	Texto	object

## Frecuencias de columnas categóricas

Columna	Valor	Frecuencia
Sexo	M	35,552
Sexo	F	30,056
Sexo	f	6
Sexo	m	2

# Módulo calidad\_datos

## Estadísticas descriptivas

Variable	Conteo	Media	Desviación estándar	Valor mín	25%	50%	75%	Valor máx
<b>Edad</b>	65,616	39.18	18.66	0.00	26.00	36.00	52.00	104.00
<b>Edad 2</b>	65,616	39.18	18.66	0.00	26.00	36.00	52.00	104.00
<b>Edad meses</b>	65,616	476.10	223.96	1.00	318.00	443.00	626.00	1,250.00

# Módulo datos\_gov

Conexión a metadatos del Portal de Datos Abiertos

1

Obtención de metadatos

2

Búsqueda de conjuntos de datos

3

Descarga de datos



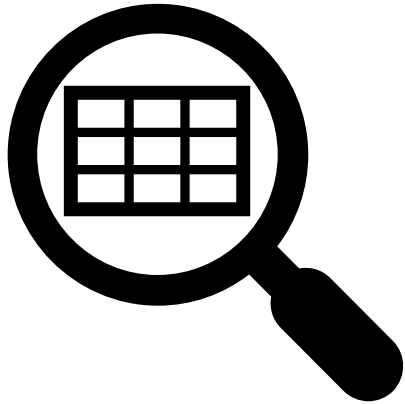
# Módulo datos Gov

Metadatos del portal datos.gov.co

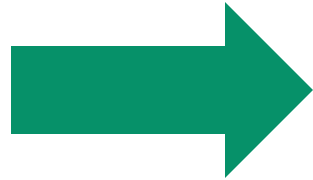


# Módulo datos\_gov

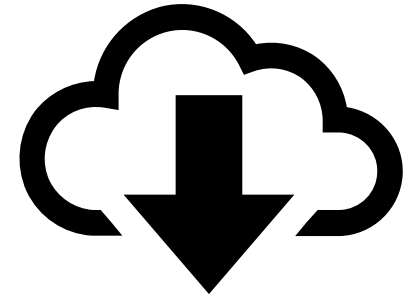
Búsqueda y descarga de conjuntos de datos



Búsqueda por términos  
clave



Filtro de conjuntos de  
datos de interés



Descarga de conjuntos  
de datos a Python

# Módulo reporte

Generación de reporte automático



# Librería LEILA

Permite generar un reporte con las métricas de calidad de la librería en un archivo HTML

## Reporte perfilamiento

Reporte generado automáticamente 27-07-2020 11:41:45 AM

### Metadatos - [ver en Datos Abiertos](#)

Atributo	Valor
<b>Id api</b>	gt2j-8ykr
<b>Nombre</b>	Casos positivos de COVID-19 en Colombia
<b>Descripción</b>	<p>Consulte los datasets históricos en: <a href="https://www.ins.gov.co/Paginas/Boletines-casos-COVID-19-Colombia.aspx">https://www.ins.gov.co/Paginas/Boletines-casos-COVID-19-Colombia.aspx</a></p> <p>ACTUALIZACIÓN: Se incluye la variable de nombre del grupo étnico. Se actualizará cada semana.</p> <p>RESUMEN DIARIO: <a href="https://infogram.com/panorama-general-1h7z2lgn3l9l4ow?live">https://infogram.com/panorama-general-1h7z2lgn3l9l4ow?live</a> Debido a los ataques que ha recibido la página del INS desde IPs extranjeras, se ha limitado el acceso de manera temporal desde IPs fuera de Colombia.</p> <p>Cualquier actualización que se identifique, quedará registrada al día siguiente en la publicación. Consulte la fe de erratas y notas aclaratorias en: <a href="http://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx">http://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx</a></p> <p>Contiene el consolidado de los casos positivos de Coronavirus COVID-19 en Colombia reportados por el Instituto Nacional de Salud (INS). Incluye variables como género, departamento, grupo étnico, entre otras.</p> <p>Para las ciudades que son distritos (Cartagena, Bogotá, Santa Marta, Buenaventura y Barranquilla), sus cifras son independientes a las cifras del departamento al cual pertenecen, en concordancia con la división oficial de Colombia.</p> <p>* Los casos marcados como en estudio están sujetos a modificación una vez se identifique el origen (importado o relacionado).</p> <p>** Recuperado es paciente con segunda prueba negativa para el virus. El paciente puede permanecer en el hospital por otras razones.</p> <p>***Por seguridad de las personas, algunos datos serán limitados evitando así la exposición y posible identificación en determinados municipios.</p>

Atributo	Valor
<b>Propietario</b>	Instituto Nacional de Salud
<b>Tipo</b>	conjunto de datos
<b>Categoría</b>	Salud y Protección Social
<b>Términos clave</b>	covid-19, coronavirus, salud, pandemia
<b>Página web</b>	<a href="https://www.datos.gov.co/d/gt2j-8ykr">https://www.datos.gov.co/d/gt2j-8ykr</a>
<b>Fecha de creación</b>	2020-03-27
<b>Fecha de actualización</b>	2020-07-26
<b>Frecuencia de actualización</b>	Diaria
<b>Número de filas</b>	240795
<b>Número de columnas</b>	21
<b>Entidad</b>	Instituto Nacional de Salud
<b>Dependencia entidad</b>	Instituto Nacional de Salud
<b>Sector entidad</b>	Salud y Protección Social
<b>Página web de la entidad</b>	<a href="https://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx">https://www.ins.gov.co/Noticias/Paginas/Coronavirus.aspx</a>
<b>Correo de contacto</b>	ccruzr@ins.gov.co
<b>Licencia</b>	Creative Commons Attribution   Share Alike 4.0 International
<b>Departamento entidad</b>	Bogotá D.C.
<b>Municipio entidad</b>	Bogotá D.C.
<b>Orden entidad</b>	Nacional
<b>Idioma</b>	Español
<b>Cobertura</b>	Nacional
<b>¿Es pública la base?</b>	SI

¿Qué incluye?

Estadísticas generales

Muestra de datos

Estadísticas específicas

Correlaciones



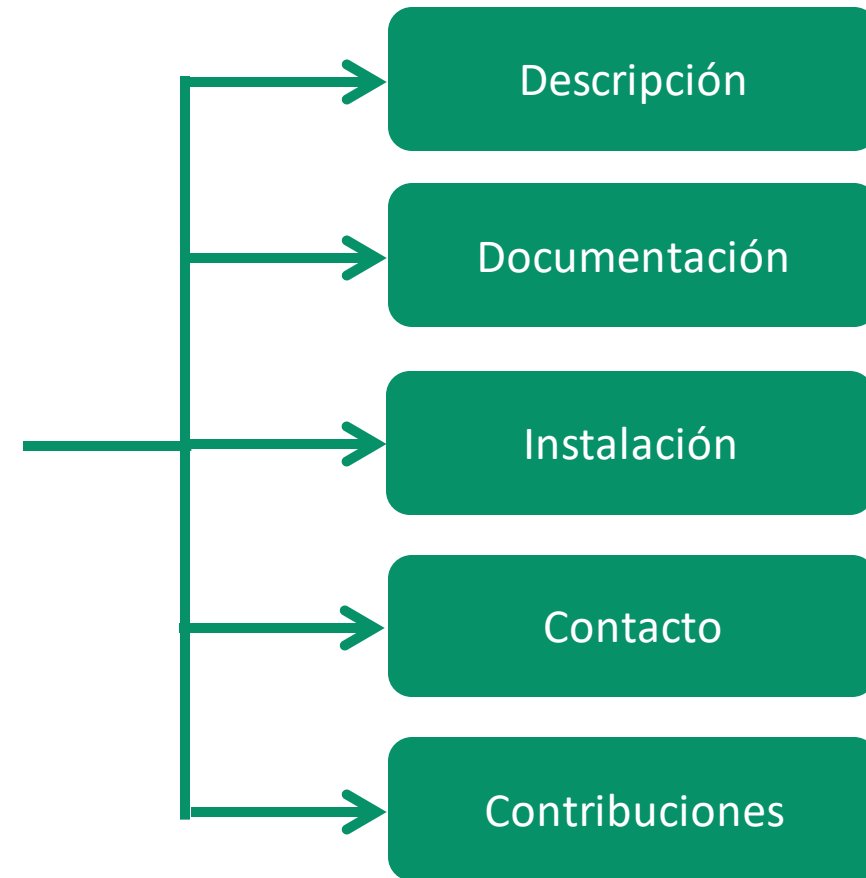


# 5. Repositorio público de GitHub

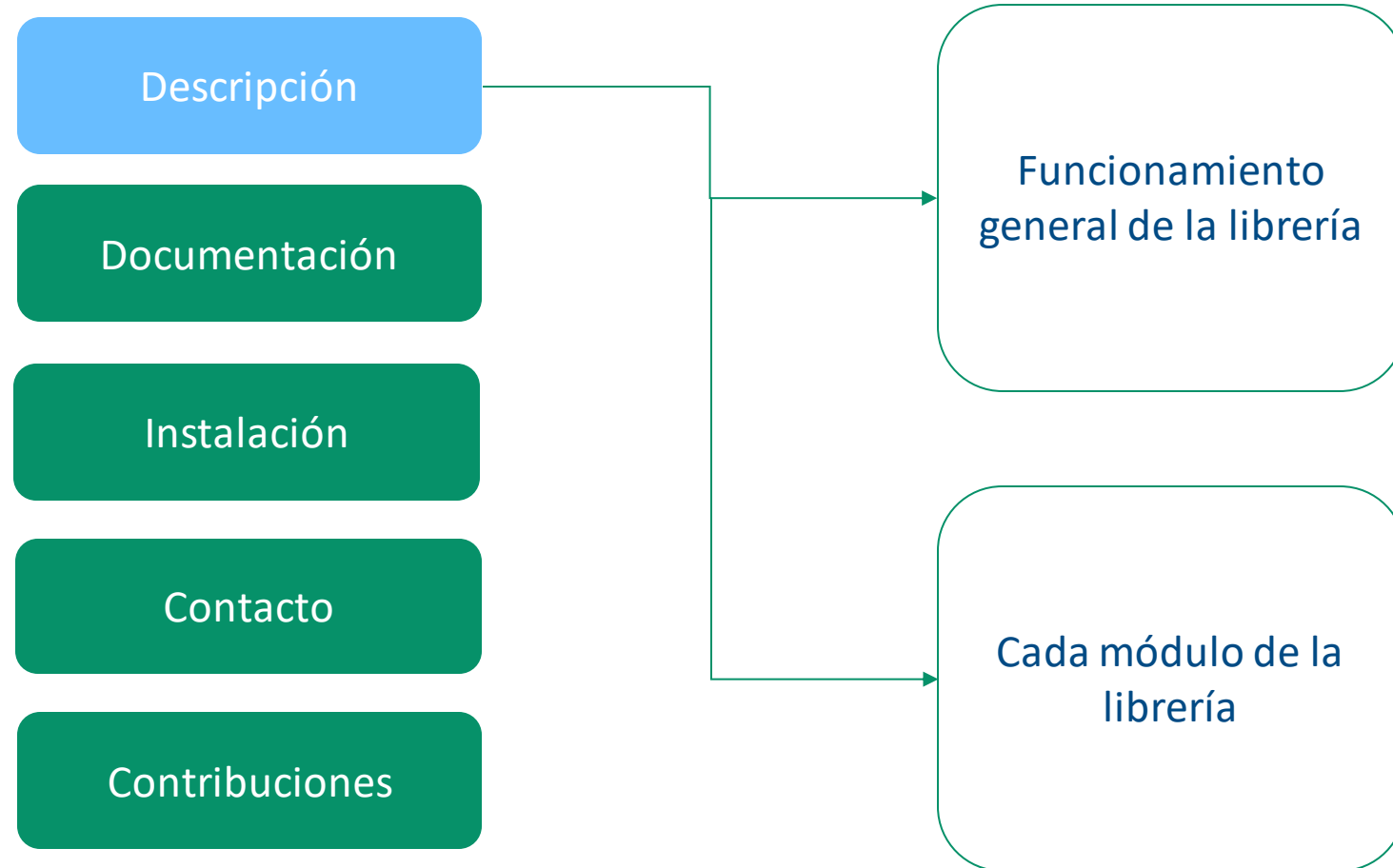
# Repositorio público de GitHub



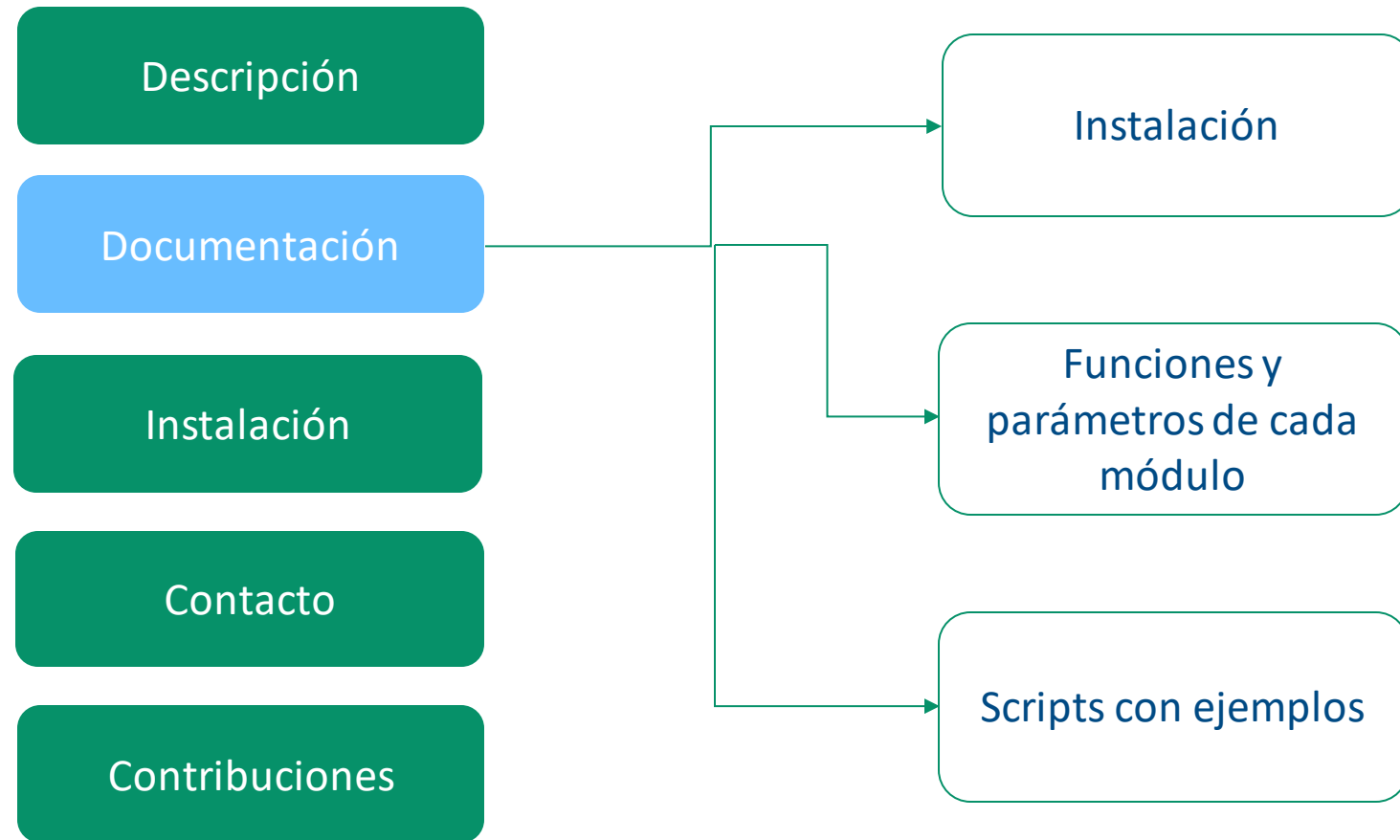
<https://github.com/ucd-dnp/leila>



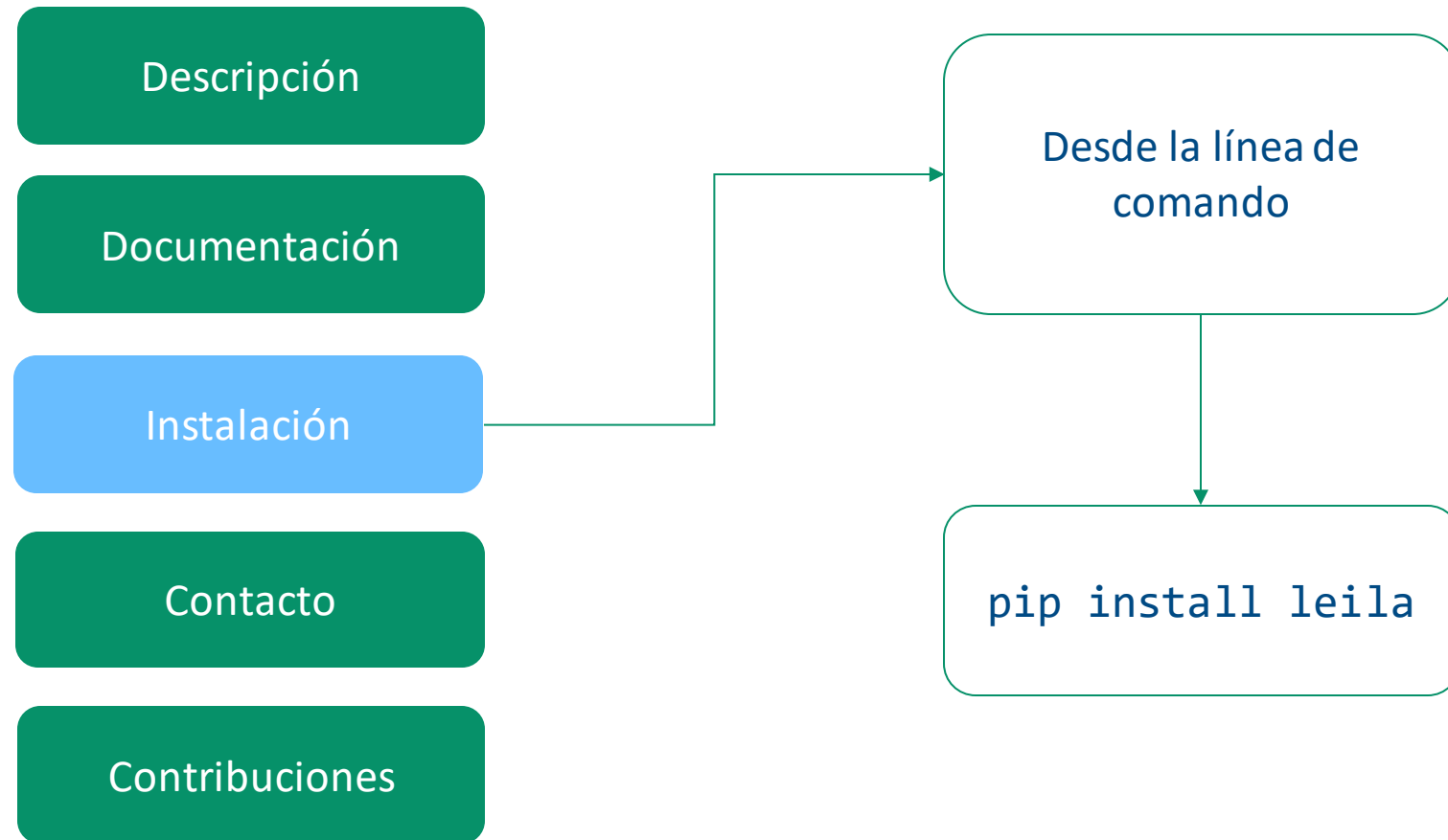
# Repositorio público de GitHub de LEILA



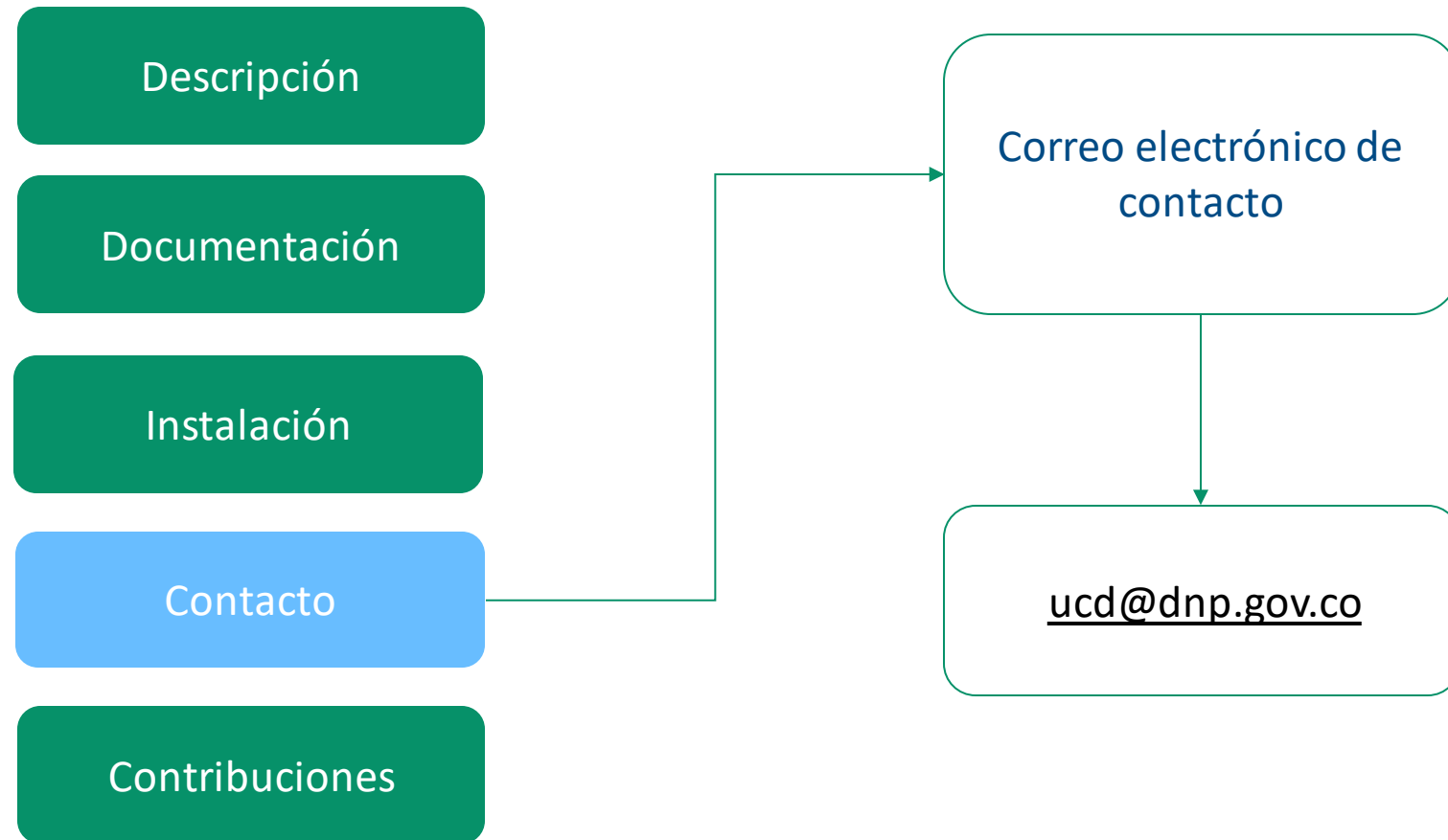
# Repositorio público de GitHub de LEILA



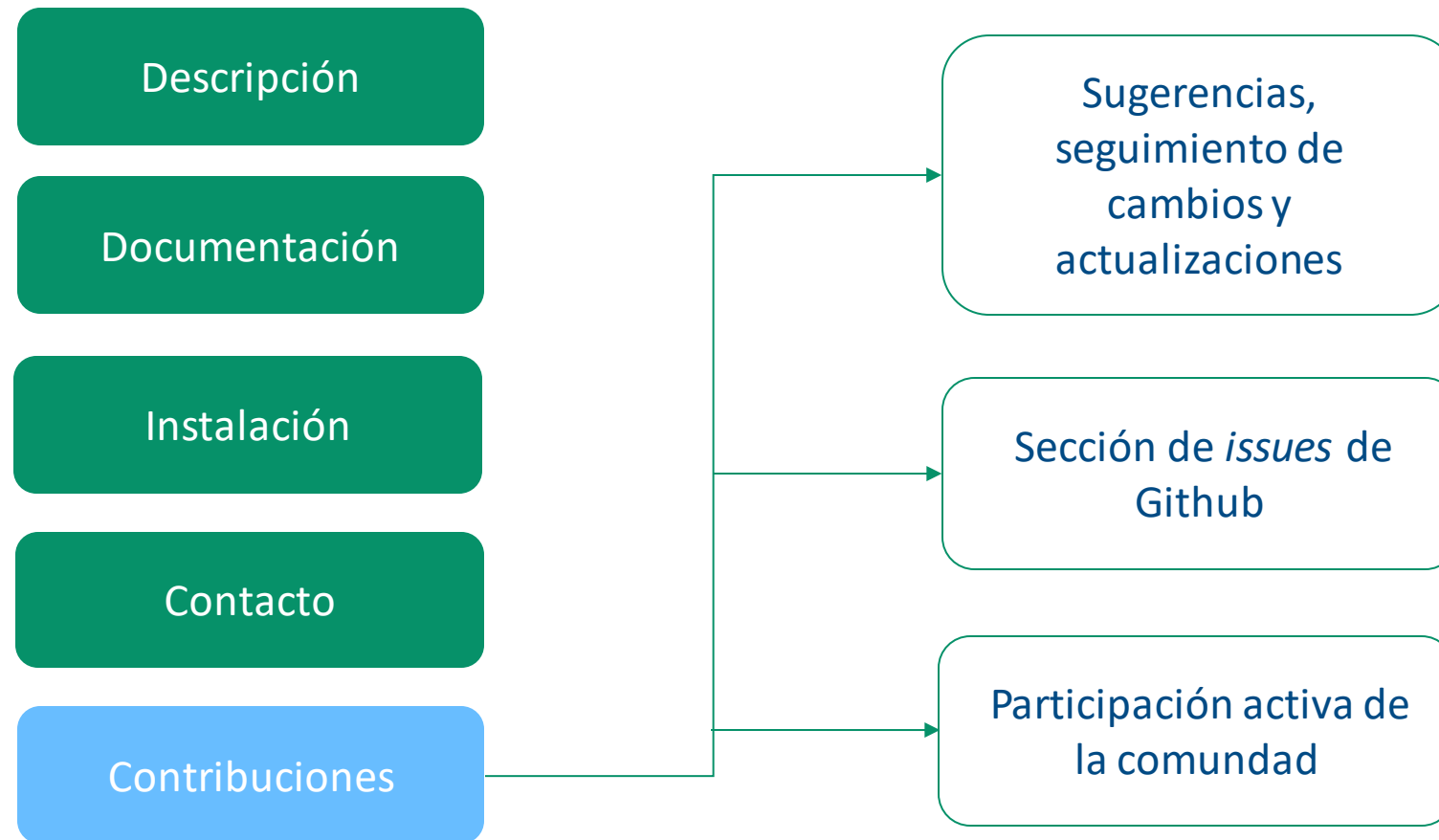
# Repositorio público de GitHub de LEILA



# Repositorio público de GitHub de LEILA



# Repositorio público de GitHub de LEILA



# 6. Consideraciones finales





# Consideraciones finales



LEILA facilita el análisis de calidad de conjuntos de datos estructurados

LEILA es de acceso público:  
libre y gratuita para todos

LEILA está abierta a comentarios y contribuciones de la comunidad

# Rutas web de interés

Ruta de GitHub de LEILA



<https://github.com/ucd-dnp/leila>

Correo de contacto UCD



[ucd@dnp.gov.co](mailto:ucd@dnp.gov.co)

Página web UCD



<https://www.dnp.gov.co/>



Direcciones



Desarrollo  
Digital



Analítica  
de Datos





**El futuro  
es de todos**

**DNP**  
Departamento  
Nacional de Planeación