Alianza SURDATA

CICLO DE TALLERES EN CIENCIA ABIERTA

TALLER 1 DE 3

Fundamentos de Datos Abiertos

Qué son los datos de calidad, qué los hace abiertos y por qué importan para el sector público, la academia y la industria.



EXPONEN

Max Koelbl
Álvaro Paredes

14 de julio · 2,5 h

DATA OBSERVATORY

Universidad de los Andes

SOBRE

La Alianza SURDATA

La **Alianza de Datos Abiertos y Responsables desde el Sur Global (SURDATA)** es una red de colaboración, que reúne a instituciones públicas, privadas, académicas y de la sociedad civil de América Latina.

Su propósito es impulsar el valor estratégico y el uso de los datos abiertos e interoperables, promoviendo principios como FAIR para fortalecer el intercambio de conocimiento y contribuir a la ciencia, la innovación y el desarrollo sostenible en la región.



PROPÓSITO

¿Qué busca la Alianza SURDATA?

Promover
la apertura, interoperabilidad y uso responsable de los datos.

Fortalecer
las capacidades humanas e institucionales mediante formación y documentación compartida.

Aplicar y enseñar
principios FAIR, con énfasis en su implementación responsable en contextos locales.

Fomentar
la colaboración como eje para diseñar políticas, estándares y herramientas comunes.

BIENVENIDA

Quiénes hacemos este taller

ORGANIZAN DATA OBSERVATORY | Universidad de los Andes

en el marco de la Alianza SURDATA

Álvaro Paredes

Director de Tecnología Data Observatory

Max Koelbl

Data Steward Dirección de Biblioteca, Universidad de los Andes

— Iremos alternando perspectivas durante todo el taller

EL RECORRIDO

Un ciclo de tres talleres

TALLER 1 HOY

Fundamentos de datos abiertos

14 de julio · 2,5 h Nivel básico · introductorio

TALLER 2

Gobierno de datos

Fecha por definir · 3 h Nivel básico, profundizando

TALLER 3

Reproducibilidad y buenas prácticas

Fecha por definir. Nivel intermedio

HOY

Objetivo y recorrido

OBJETIVO DEL TALLER

Entender qué son los datos de calidad, los datos abiertos y su beneficio en distintos ámbitos.

01 ¿Qué entendemos por datos?

02 Por qué importan los datos

03 Datos abiertos y principios FAIR

04 Lo que hace reutilizables a los datos

05 Confianza: calidad, gobernanza y ética

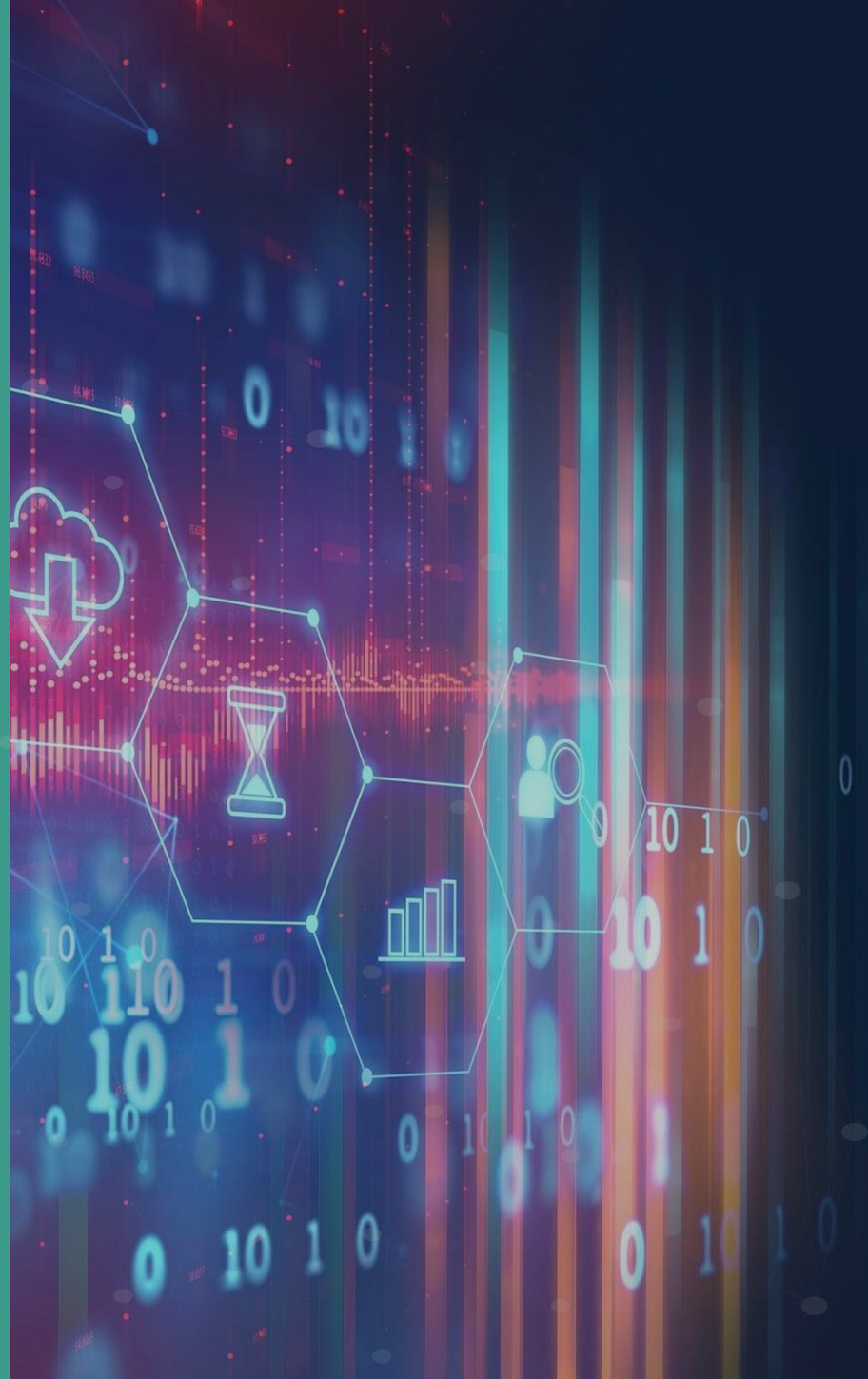
\+ Discusión y preguntas

01

BLOQUE 01

¿Qué entendemos por datos?

Una base común antes de hablar de apertura: qué es un dato, cómo vive y por qué los metadatos importan tanto como el dato mismo.



HILO CONDUCTOR

Un ejemplo que nos acompaña todo el día

Cada concepto lo aterrizamos sobre un caso cotidiano, para que el beneficio se vea claro.

EJEMPLO 1

Transporte público

Paraderos, recorridos y tiempos de llegada de los buses.

Lo usamos a diario y lo tocan los tres sectores: el Estado lo opera, la academia lo estudia y empresas crean apps con estos datos.

EJEMPLO 2

Soledad y aislamiento durante la pandemia

Proyecto interdisciplinar de salud mental, con apoyo de ministerios, prestadores de servicios del ámbito de salud y organizaciones comunitarias.

Ideas: un dataset de investigación en un repositorio, o un registro bibliográfico. A definir y enlazar a cada concepto.

CONCEPTO BASE

¿Qué es un dato?

El registro mínimo de un hecho, un número, un texto, una fecha. Por sí solo dice poco.

Su valor aparece cuando lo conectamos, lo interpretamos y lo usamos para decidir. Ese camino es el que queremos cuidar.

DATO	"18:42" · "Bus 506"
INFORMACIÓN	El bus 506 llega al paradero a las 18:42
CONOCIMIENTO	En hora punta, el 506 se atrasa ~8 min
DECISIÓN	**Reforzar la frecuencia a las 18:00**

PARA PONER EN CONTEXTO

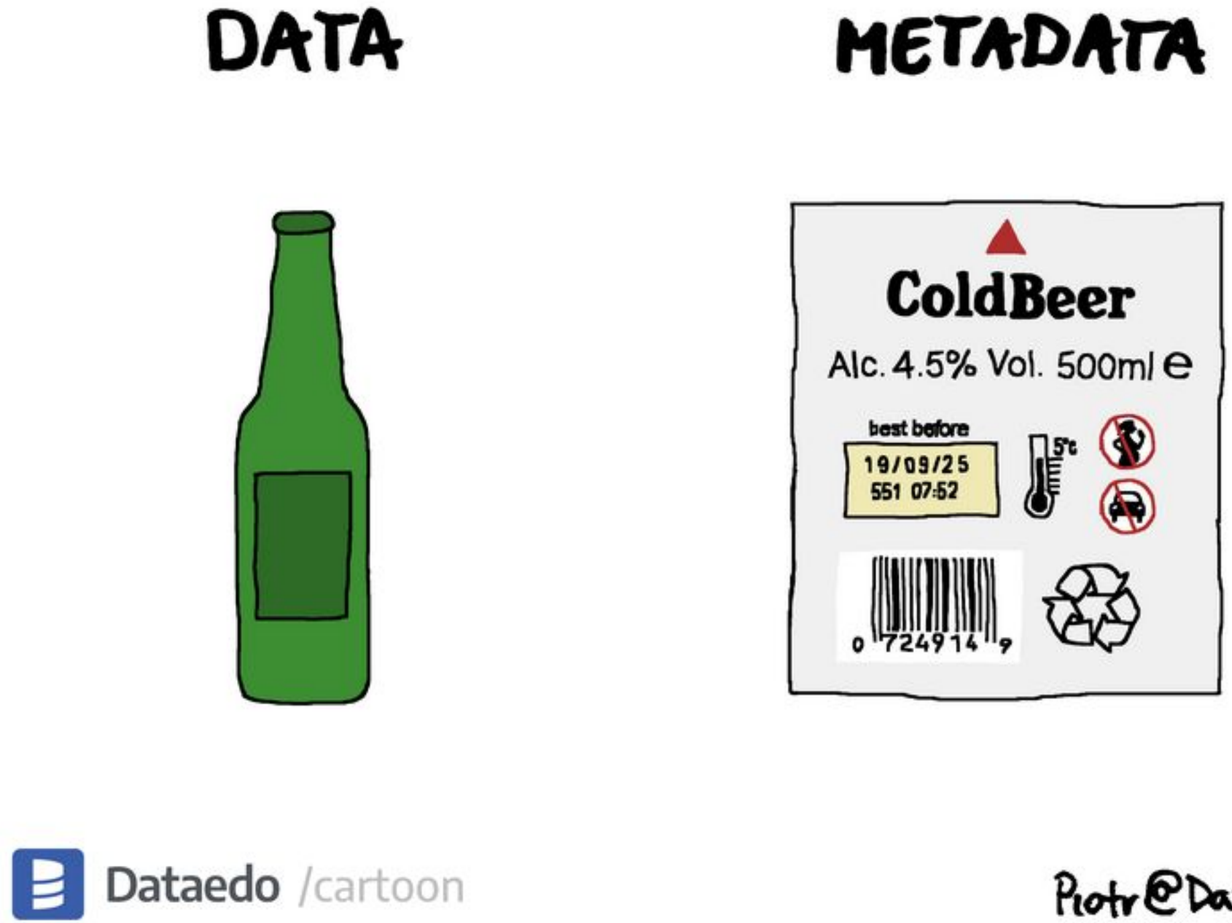
El ciclo de vida de los datos

1	**Capturar**	Sensores en buses y paraderos registran tiempos y posiciones.
2	**Procesar**	Limpiar, estructurar y validar los datos en bruto.
3	**Analizar**	Detectar patrones: horarios críticos, líneas con mayor atraso.
4	**Compartir**	Publicar y abrir el dataset con licencia y metadatos.
5	**Reutilizar**	Nuevos usos y valor: apps, investigación, decisiones de política.

↻ **Preservar** cruza todo el ciclo: mantener los datos accesibles y entendibles en el tiempo.

DISTINCIÓN CLAVE

Dato y metadato



DISTINCIÓN CLAVE

Dato y metadato

El dato es el contenido. El **metadato es el dato sobre el dato** : el contexto que lo hace encontrable, entendible y reutilizable.

EL DATO

Llegadas del bus 506: 18:42 · 18:55 · 19:09

EL METADATO

Fuente: **DTPM**

Cobertura: **RM**

Actualizado: **hoy 06:00**

Formato: **CSV**

Unidad: **hora local**

Licencia: **CC BY 4.0**

CUATRO TIPOS DE METADATO

Descriptiva

Autor, título, tópicos, fecha, ubicación. Para buscar e identificar.

Técnica

Formato, estructura, codificación, ubicación del archivo.

Legal

Licencia, copyright y propietario de los datos.

Procedencia

Origen, métricas y proceso con que se generó el recurso.

02

BLOQUE 02

Por qué importan los datos

El beneficio concreto: economía de los datos, decisiones con evidencia, inteligencia artificial y transformación digital en los tres sectores.



EL VALOR

La economía de los datos

Los datos no se comportan como otros recursos: no se gastan al usarse y valen más cuando se comparten.

No se agotan
El mismo dato puede usarse muchas veces, por muchos, a la vez.

Se revalorizan al combinarse
Transporte + clima + demografía revelan más que cada uno por separado.

Habilitan nuevos servicios
Apps de rutas, planificación urbana, investigación: todo sobre el mismo dato abierto.

EN LA PRÁCTICA

Decisiones basadas en datos

SIN DATOS

"Reforcemos la frecuencia donde **creemos** que hace falta."

CON DATOS

"Reforcemos donde se cruzan **mayor atraso** y **mayor demanda** ."

Datos → **Evidencia** → **Decisión** → **Resultado** ↺

La IA necesita datos de calidad

Un modelo es tan bueno como los datos que lo alimentan.

Si los tiempos de llegada están incompletos o desactualizados, la predicción del próximo bus también lo estará.

"Basura entra, basura sale."

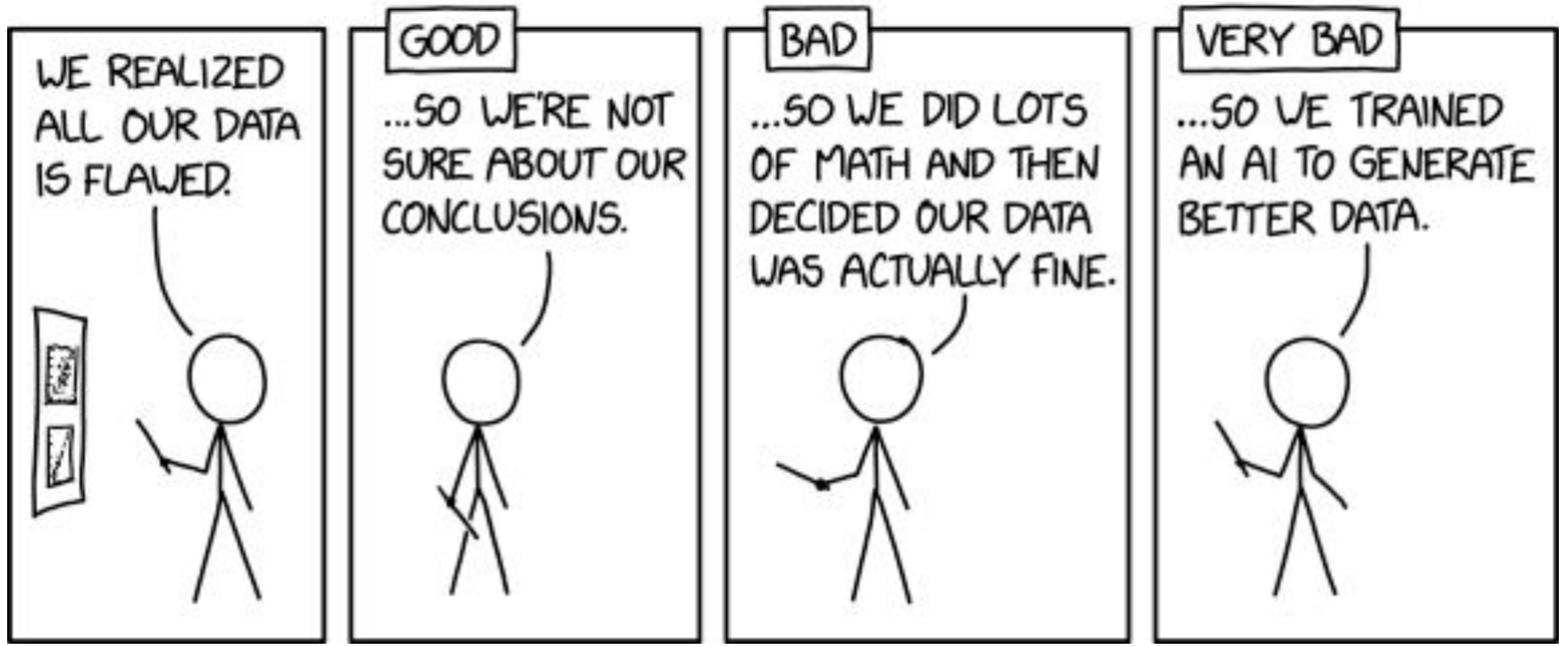
Volumen
Suficientes ejemplos para aprender patrones.

Calidad
Correctos, completos y actualizados.

Representatividad
Datos sesgados producen resultados sesgados.

Hasta el mejor modelo falla con datos malos

Datos de calidad son el insumo básico de toda innovación.



https://xkcd.com/2494/

BENEFICIO DIRECTO

¿Por qué invertir en datos abiertos?

Transformación digital en tres sectores

Gobierno

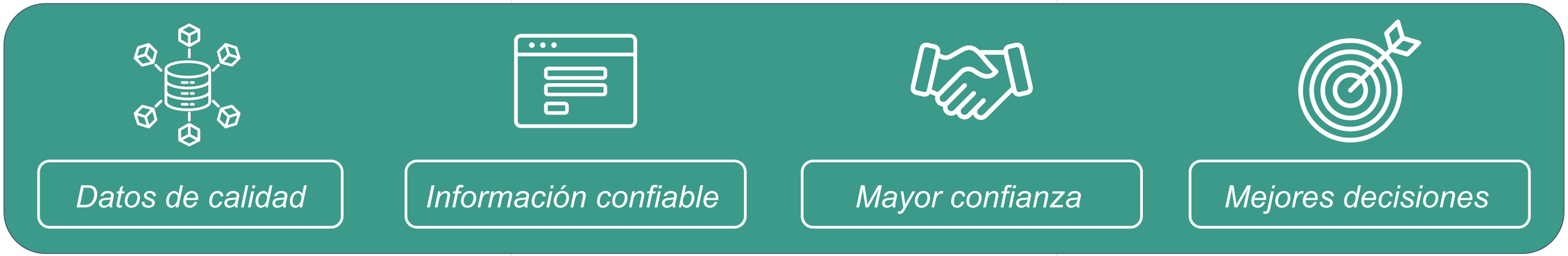
Mejores servicios y políticas públicas, asignación más eficiente de recursos, más transparencia y rendición de cuentas.

Academia

Investigación reproducible, colaboración y reutilización, generación de nuevo conocimiento y más visibilidad.

Sector privado

Mayor eficiencia operativa, menos riesgos y errores, innovación, nuevos productos y servicios construidos sobre datos abiertos.



→ Planifica mejor el transporte.

→ Estudia movilidad y equidad urbana.

→ Crea apps de rutas y llegadas.

BENEFICIO DIRECTO

¿Por qué hacerlo?

El ejemplo del software abierto

- The Value of Open Source Software (Harvard Business School, 2024)
  - Manuel Hoffmanna, Frank Nagleb, Yanuo Zhou
- $4,15B invertidos en Código abierto, genera $8,8T de valor para compañías ($1 -> $2.000)
- Compañías necesitarían invertir 3,5 veces más en software de lo que lo hacen actualmente, si OSS no existiera.
- Este documento habla de Software, pero podemos pensar en una sinergia similar con (meta)datos.

03

BLOQUE 03

Datos abiertos y principios FAIR

Qué hace que un dato sea verdaderamente abierto y, sobre todo, en qué se diferencia de ser FAIR.



DEFINICIÓN

¿Qué son los datos abiertos?

Datos que **cualquiera puede usar, reutilizar y redistribuir libremente**, a lo más con la condición de citar la fuente.

CERRADOS	COMPARTIDOS	ABIERTOS
Uso interno. Sin acceso ni reutilización externa.	Acceso a grupos definidos, con condiciones y control.	Cualquiera puede usar, reutilizar y redistribuir libremente.
		Formato adecuado + licencia clara.

> *"As open as possible, as restricted as necessary."*

Abrir no es solo subir un archivo: importan el formato y la licencia



Domingo 21 junio de 2026 | 14:00

Roban notebook con años de estudios sobre bacterias marinas a investigadora de la USM en Viña del Mar

Alberto González
Periodista de Prensa en BioBioChile

Con información de Benjamín Duque

Seguimos criterios de The Trust Project

Ética y transparencia de BioBioChile



ARCHIVO | Agencia UNO | USM

visitas

VER RESUMEN

Resumen generado con una herramienta de Inteligencia Artificial desarrollada por BioBioChile y revisado por el autor de este artículo.

Computador con años de investigación científica sobre bacterias robado en Viña del Mar. Doctora en microbiología y académica de la USM, Fernanda Claverías, víctima del ilícito. Equipo sustraído desde vehículo en 7 Norte con 1 Poniente. Avalúo inicial de $500 mil, pero real valor en valiosa data científica. Cámaras registran robo, diligencias en curso. Estudios de Claverías apuntaban a bacterias marinas con potencial antibiótico en Valpo

Suscríbete a nuestros Newsletters

Recibe lo mejor de BioBioChile en nuestros Newsletters, entregados directamente a usted de forma gratuita

Ingresa aquí

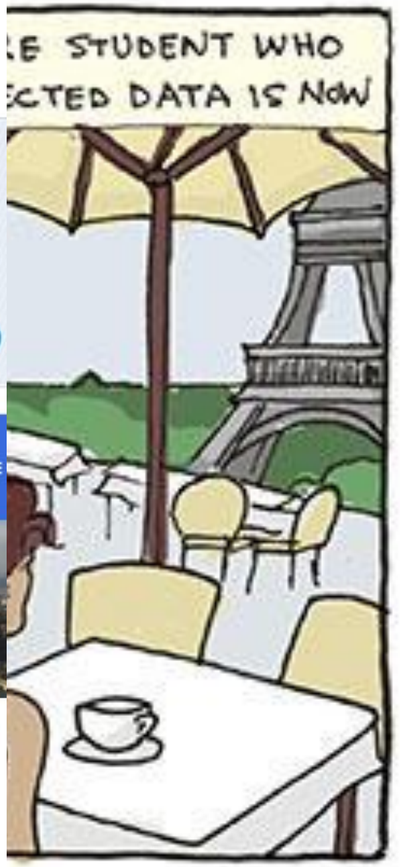
cchc GENTE QUE CONSTRUYE



80% de las manzanas de Santiago tiene un solo uso de suelo: CChC releva impacto de la mixtura urbana

- Expertos destacan impacto del urbanismo en la calidad de vida: anuncian detalles de encuentro CIC 2026
- Nuevo subsidio para viviendas de hasta 4.000 UF: revisa quiénes podrán acceder a este beneficio

MÁS NOTICIAS



EL ESTÁNDAR

Los principios FAIR

F

Findable · Localizable

Fácil de encontrar: identificador único, metadatos descriptivos e indexado en recursos buscables.

A

Accessible · Accesible

Obtenible mediante protocolos estándar, incluso si requiere autenticación.

I

Interoperable

Usa formatos y vocabularios estándar; combinable con otros datos por máquinas.

R

Reusable · Reutilizable

Bien documentado, con licencia clara y descripción de procedencia.

FAIR son principios, no estándares rígidos — sirven a personas y máquinas por igual

LA DIFERENCIA CLAVE

Abierto no es lo mismo que FAIR

ABIERTO

Es una cuestión de **permisos** : ¿puedo usarlo y redistribuirlo libremente?

FAIR

Es una cuestión de **reutilización** : ¿puedo encontrarlo y reutilizarlo bien, humanos y máquinas?

FAIR pero no abierto

Datos de salud sensibles: metadatos públicos y acceso controlado.

Abierto pero no FAIR

Un PDF colgado sin metadatos: libre pero difícil de hallar y reutilizar.

El ideal

Abierto **y** FAIR: libre de usar y fácil de reutilizar.

04

BLOQUE 04

Lo que hace reutilizables a los datos

Las piezas concretas detrás de FAIR: licencias, identificadores persistentes y formatos.



PERMISOS DE USO

Licencias

Una licencia dice **qué se puede hacer con el dato**. Sin ella, la duda legal frena la reutilización.

CC0 — **Dominio público**
Sin condiciones. Cualquiera puede usar, modificar y redistribuir sin atribución.

CC BY — **Atribución**
Úsalo libremente, citando la fuente. La licencia más común en datos abiertos.

BY-SA — **Atribución / Compartir igual**
Cita la fuente y comparte en los mismos términos de los que se derivan estos datos.

> Ejemplo: el dataset de transporte con CC BY 4.0 ya puede alimentar cualquier app con solo citar al DTPM.

PID

Identificadores persistentes

Un identificador permanente para datos, personas e instituciones. No cambia aunque cambie la URL — los enlaces no se rompen y la atribución es inequívoca.

DOI

Digital Object Identifier

Identifica datasets y publicaciones de forma permanente. Estándar global para citar datos.

10.71481/7XPJ-CV52

ORCID

Open Researcher ID

Identifica a las personas investigadoras a lo largo de su carrera.

ROR

Research Org Registry

Identifica instituciones. ISNI cumple un rol similar en el ámbito cultural.

INTEROPERABILIDAD PRÁCTICA

Formatos: PDF, CSV y API

El formato decide cuánto se puede reutilizar un dato. Mismo contenido, muy distinto potencial.

PDF	**Para leer, no para procesar.** El horario impreso del paradero.	BAJA
CSV	**Estructurado, legible por máquinas.** Una tabla de paradas y horarios.	MEDIA
API	**Datos en vivo, consumibles por cualquier app.** Llegadas en tiempo real.	ALTA

05

BLOQUE 05

Confianza: calidad, gobernanza y ética

Para que un dato genere valor, hay que poder confiar en él y usarlo de forma responsable.



CONFIAR EN EL DATO

Calidad, integridad y actualidad

Calidad

Correctos, completos y consistentes. Sin errores ni vacíos que engañen al análisis.

Integridad

No alterados y trazables: se sabe de dónde vienen y cómo se obtuvieron.

Actualidad

Vigentes. Tiempos de llegada de hace un año no sirven para decidir hoy.

Completitud

Contiene todos los valores necesarios: se obtuvieron todos los valores por medir para un rango de tiempo definido.

CONCEPTOS CLAVE

Gobernanza, gestión y DAMA

GESTIÓN DE DATOS

El hacer del día a día

Capturar, almacenar, limpiar, mantener y publicar los datos. Las operaciones concretas.

GOBERNANZA DE DATOS

Las reglas del juego

Quién decide, quién es responsable de cada dataset y bajo qué políticas.

DAMA **Data Management Association International**

El DMBOK (Data Management Body of Knowledge) es el marco de referencia profesional. Define 11 áreas de conocimiento — arquitectura, calidad, seguridad, metadatos, entre otras. Lo usaremos como base en el **Taller 2**.

USO RESPONSABLE

Ética y datos

Abrir **no** es abrir todo. La apertura es responsable, no ingenua.

Privacidad

Proteger los datos personales y sensibles. No todo puede ni debe abrirse.

Consentimiento

Usar los datos únicamente para los fines que se autorizaron.

Sesgos

Vigilar qué o a quién dejan fuera los datos que usamos.

Transparencia

Ser claros sobre el origen, los fines y los límites de los datos.

¿Cuánto es una marraqueta?



¿cuánto es una docena? ¿una resma?

https://www.reddit.com/r/chile/comments/p77tbw/cuanto_es_1_marraqueta/

DOS MUNDOS, UNA MISIÓN

Data Stewardship académico y corporativo

Misión principal

Asegurar la reproducibilidad, la transparencia y la preservación a largo plazo de la ciencia.

Garantizar la calidad y valor del dato como activo estratégico para la toma de decisiones.

Marco de trabajo

Investigación y gestión de datos de Investigación (RDM). Ciencia abierta y principios FAIR.

Gobierno de datos corporativo (data governance program office). Reglas de negocio, implicancias comerciales y operativas.



DOS MUNDOS, UNA MISIÓN

Data Stewardship académico y corporativo

Enfoque operativo

Acompañamiento e incentivo al investigador en sus planes de gestión de datos (DMPs). Soporte específico según la disciplina .

Definición de metadatos, linaje del dato y establecimiento del estándar de oro. Resolución de duplicados, calidad del dato y definición de "data ownership".

Naturaleza de la autoridad

Facilitador y asesor técnico. El investigador o laboratorio retiene el control directo de los datos.

Autoridad formalizada y delegada. Es el referente para dirimir significados y tiene el poder de aprobar cambios y hacer cumplir decisiones sobre el dato.



DOS MUNDOS, UNA MISIÓN

Data Stewardship académico y corporativo

Destino del dato

Repositorios públicos o de largo plazo (e.g., Dataverse, Zenodo) para su reutilización global.

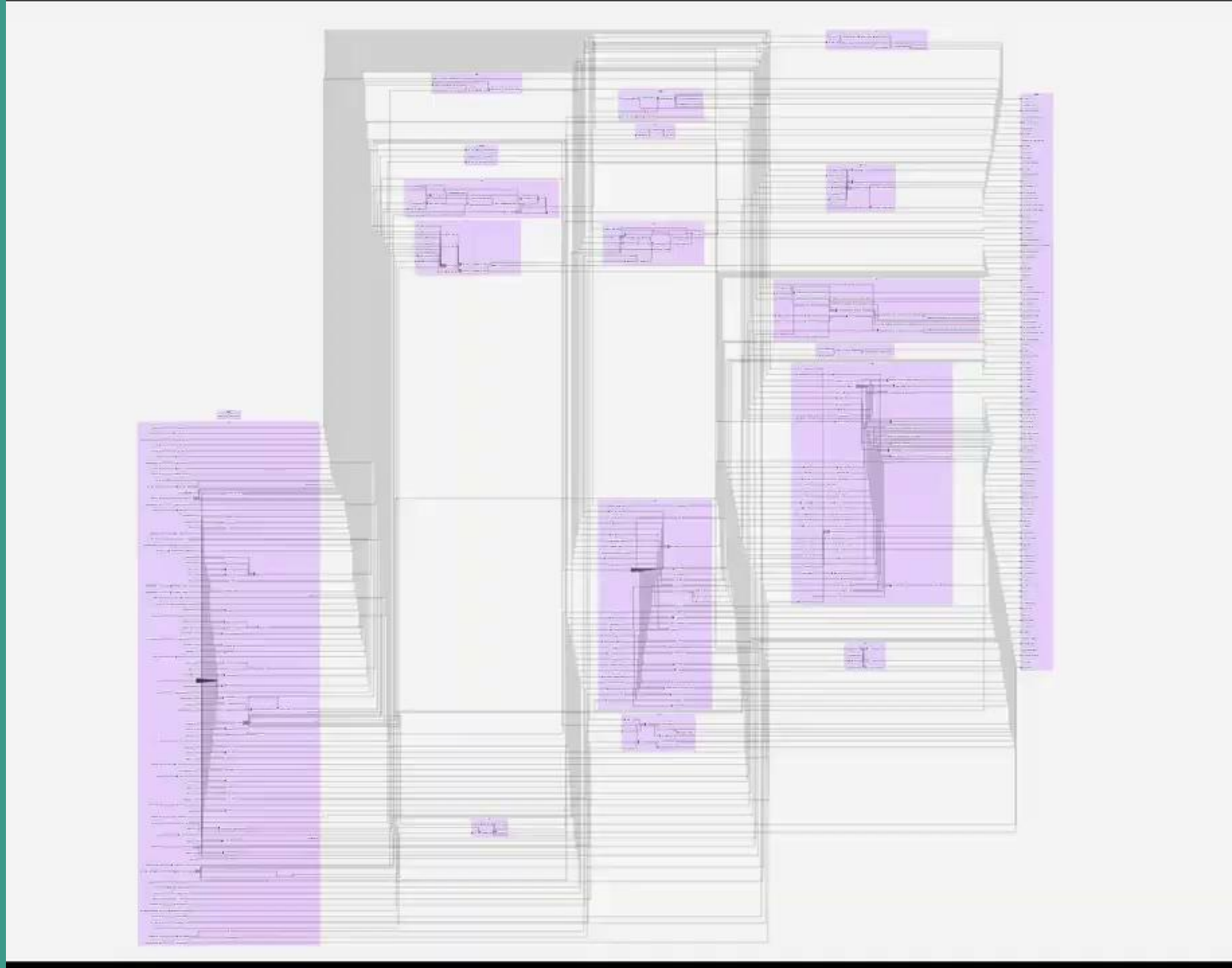
Almacenes de datos (Data Warehouses/Lakes) y sistemas transaccionales de uso interno y restringido.

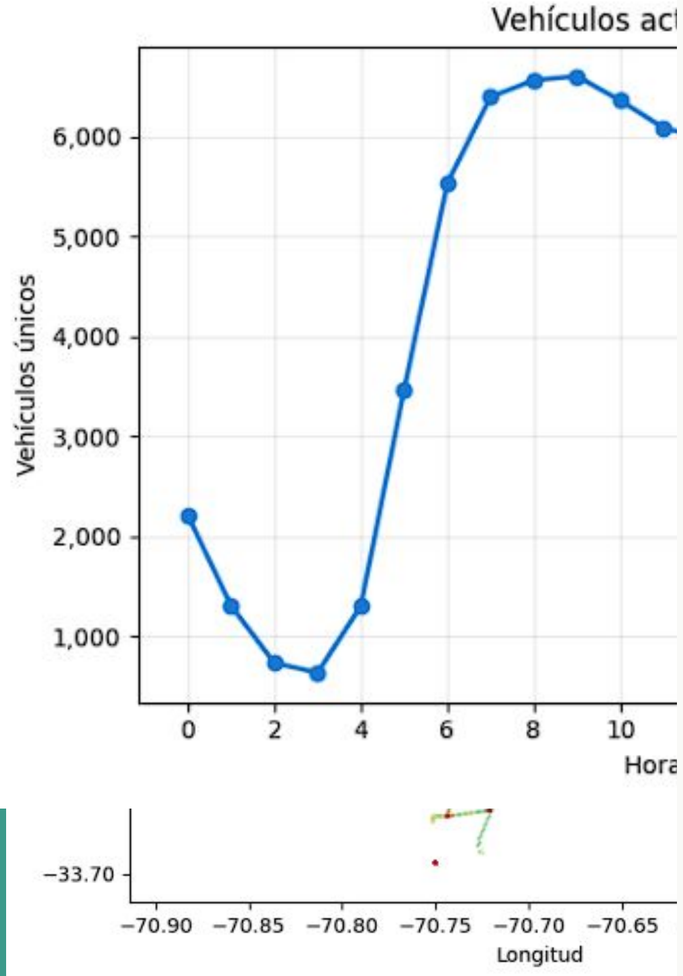


¿Hacia dónde queremos ir?

Algunos ejemplos







RED Movilidad — posiciones GPS diarias

Estructura de partición en S3 (GeoParquet) y distribución de archivos por mes · Datopia Lakehouse

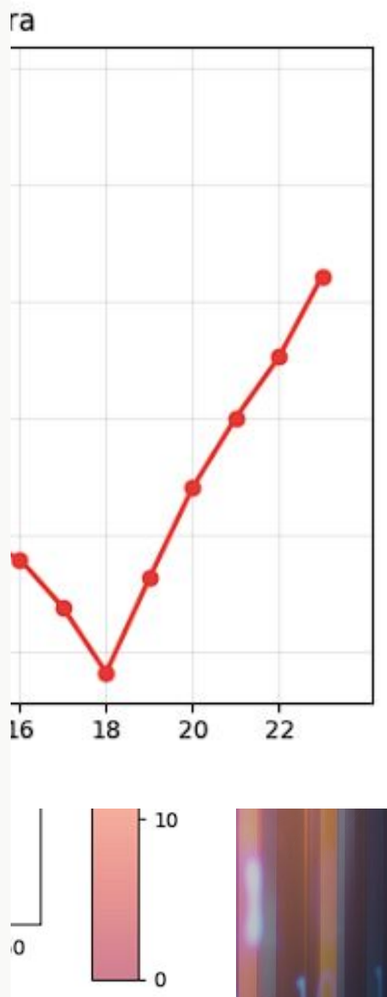
ÁRBOL DE PARTICIÓN Y METADATOS (STAC-LIKE)

```
s3://do-datopia/
├─ catalog.json                                  # Catálogo raíz · 2 categorías
└─ categoria=transporte/
   ├─ collection.json                            # Colección · 2 datasets (RED, futuros)
   └─ pais=cl/fuente=red-movilidad/tipo=posicion-diaria/version=v1/
      ├─ dataset.json                            # Dataset · esquema 13 columnas, geometría, partición
      ├─ year=2025/
      │  ├─ month=09   (22 días)
      │  ├─ month=10   (31 días)
      │  ├─ month=11   (30 días)
      │  └─ month=12   (31 días)
      └─ year=2026/
         └─ month=01 … month=07   (6 completos + jul en curso)
            └─ day=DD/                            # 301 particiones día (una por fecha UTC)
               ├─ manifest.json                   # Item · row_count, bytes, parts[], created_at_utc
               └─ part-00000.parquet              # datos GeoParquet, ~4–13M filas
```

Jerarquía de metadatos tipo **STAC**: Catalog → Collection → Dataset → Item (manifest). Cada nivel es JSON leíble directo desde S3 — sin base de datos externa.

Partición de datos: **year / month / day** · compresión **zstd** · geometría **WKB Point EPSG:4326** · orden de escritura **timestamp_gps_utc, plate**.

ARCHIVO	NIVEL	CAMPOS CLAVE	CARDINALIDAD
`catalog.json`	Catalog	`id`, `categorias[]` (id, href, dataset_count), `generated_at_utc`	**1 (raíz del bucket)**
`collection.json`	Collection	`categoria`, `datasets[]` (id, fuente, tipo, version, formato, href), `generated_at_utc`	**1 por categoría (2)**
`dataset.json`	Dataset	`description`, `format`, `partitioned_by`, `columns[]` (name, type, comment), `geometry`, `sort_order`, `schema_version`	**1 por dataset (1)**
`manifest.json`	Item	`date_utc`, `row_count`, `part_count`, `parts[]` (file, row_count, bytes), `created_at_utc`, `generator`	**1 por día (301)**



RECAPITULAMOS

Lo que nos llevamos

01 Un dato vale cuando se **conecta e interpreta** para decidir.

02 El **metadato** hace usable al dato — cuatro tipos.

03 **Abierto ≠ FAIR**: permisos frente a reutilización.

04 **Licencias, PIDs y formatos** son los habilitadores concretos.

05 **Calidad, gobernanza y ética** son la base de la confianza.

06 El beneficio es **directo para los tres sectores**.

EL RECORRIDO CONTINÚA

Qué sigue

TALLER 2

Fecha por definir · 3 h

Gobierno de Datos

Profundizar en FAIR y traducir a decisiones organizacionales: roles, responsabilidades y los mínimos que necesita un equipo. Profundizamos DAMA.

TALLER 3

Fecha por definir

Reproducibilidad y Buenas Prácticas

Transformar un conjunto de datos en un flujo de trabajo claro, reproducible y compartible. Enfoque práctico.

CONVERSEMOS

Discusión y preguntas

¿Qué datos abren o usan hoy en su institución?

¿Qué los frena para abrirlos o reutilizarlos?

¿Dónde podrían empezar esta semana?

Alianza SURDATA

CICLO DE TALLERES EN CIENCIA ABIERTA

Gracias

Nos vemos en el Taller 2 · **Gobierno de Datos** (fecha por definir).



EXPONEN

Álvaro Paredes · Max Koelbl

CONTACTO

contacto@dataobservatory.net

DATA OBSERVATORY