Catálogo de datos

Table of contents

Los datasets expuestos por la API viven en `data/meta/catalog/` (YAML/JSON). Cada archivo define metadatos, orígenes, capas y linaje.

1 Estructura base

```yaml
slug: epdatos
nombre: Base de datos transversal
descripcion: Text corto...
origen:
  fuente: Espacio Público
```

```yaml
  url: https://ep.illanes00.cl
  periodicidad: mensual
capas:
  raw:
    archivos: []
    notas: ...
  bronze:
    tabla: bronze.pending
    transformaciones: []
  silver:
    tabla: silver.pending
    transformaciones: []
  gold:
    tabla: gold.pending
    transformaciones: []
dominios:
  - slug: geografia
    nombre: Geografía y territorio
esquema:
  tablas:
    - nombre: dim_region
      campos:
        - nombre: region_id
          tipo: smallint
lineaje:
  raw: []
  bronze: []
  silver: []
  gold: []
contacto:
  nombre: Equipo Datos Espacio Público
  email: datos@espaciopublico.cl
```

2 Convenciones

- `slug` debe coincidir con la ruta API (`/datasets/{slug}`).
- `dominios` agrupa temas y subdominios.
- `esquema` lista tablas y campos publicados (reflejo de Postgres).
- `lineaje` describe transformaciones entre capas raw → gold.
- `contacto` identifica responsable operativo.

3 Validación

1. Añade el archivo YAML y ejecuta `make lint` (valida formato con `pre-commit`).

2. Revisa la respuesta `GET /api/0.6.9/datasets/{slug}`.
3. Documenta el dataset en `docs/api_catalog.md` si introduce endpoints adicionales.
4. Para exploración tabular, utiliza `/api/{version}/data/{schema.table}` (filtros, paginación y linaje incluidos).

4 Instrumentación

- Los catálogos son cacheados por la API (`@cached_response`).
- Métricas se registran en `/api/0.6.9/status` con `payload_meta.type = dataset`.

5 Próximos pasos

- Incorporar validadores automáticos (pydantic) sobre los YAML.
- Sincronizar catálogos con `data/meta/lineage/`.
- Generar documentación estática a partir de estos metadatos (pages/pdfs).

6 DIPRES (Raw)

- `data/meta/catalog/dipres_ejecucion_total.yaml`: inventario de Ejecución Total.
- `data/meta/catalog/dipres_proyecto_ley.yaml`: Proyecto de Ley de Presupuestos (nivel nacional).
- `data/meta/catalog/dipres_ley_presupuesto.yaml`: Ley vigente (nacional / partidas legacy).
- Manifest consolidado: `data/meta/dipres_files_metadata.csv` accesible vía `/downloads/data/meta/dipr`
- API de descargas: `/api/{version}/datasets/{slug}/raw-assets` desde la landing o vía cliente HTTP.
- Ver dipres_catalog.md para el detalle de scraping, cobertura y backlog.
- Los archivos fuera de `data/` se replican automáticamente en `data/cache/raw_assets/` para garantizar descargas desde `/downloads/*`.

7 DIPRES (Silver)

- `data/meta/catalog/dipres_presupuestos.yaml`: catálogo del presupuesto normalizado.
- `data/meta/schema/dipres_presupuestos.sql`: definición de tablas (presupuestos, partidas, capítulos, etc.).
- `data/meta/diccionarios/dipres_subtitulos.yaml`: glosario por subtítulo.
- Expuesto vía `/api/{version}/presupuestos` (ver documento de API).

8 DIPRES (Referencias complementarias)

- `data/meta/catalog/dipres_anexo4.yaml`: catálogo específico del Anexo 4 con visor `/catalogo/dipres_anexo4`.
- `data/ref/dipres/anexo4_classificacion.yaml`: mapeo jerárquico partida → capítulo → programa → subtítulo → ítem → asignación según Anexo 4 del informe DIPRES. Incluye inclusiones y exclusiones documentadas en el PDF.
- `data/ref/dipres/cuadro_ii_2_reclasificacion.csv`: tabla manualmente estructurada del Cuadro II.2 con la reclasificación de subtítulos y notas de consolidación "bajo la línea".
- `data/ref/dipres/indicadores_I6.csv`: extracto del apartado I.6 con IPC (base 2023=100), PIB nominal, tipo de cambio promedio y de cierre.
- `data/meta/catalog/dipres_estadisticas_cofog.yaml`: series COFOG oficiales (Gobierno Central Total, Presupuestario y con Empresas Públicas) y clasificadores funcionales niveles 1-3. Fuente: excels `GCT/GCP/GCEP/Funcional`.
- `data/meta/catalog/dipres_estadisticas_clasificacion.yaml`: clasificación económica del Estado de Operaciones (Gobierno General y Municipal) en pesos corrientes, reales (base 2024) y % del PIB.
- `data/meta/catalog/dipres_recuadros_art5517.yaml`: recuadros trimestrales del Informe de Finanzas Públicas (series GC, GG y Municipal), disponibles en los excels "Primer/Segundo/Tercer/Cuarto informe".
- QA adicionales: (i) sumatoria COFOG por año vs. "Gasto Total" del Excel (<0,1 % de brecha permitida); (ii) clasificación económica corriente vs. real – verificación de deflactor relativo; (iii) % del PIB recalculado contra serie PIB Dipres (recuadro Art. 5517).
- QA aplicable:
  - Anexo 4: validación de unicidad `{partida,capítulo,programa}` y consistencia frente a catálogos internos.
  - Cuadro II.2: revisión manual (pendiente automatizar parseo directo desde PDF).
  - Indicadores I.6: comparados contra series BCCh (IPC promedio anual rebased a 2023=100, brecha <0.25 pts).
- Visualizadores HTML disponibles en `/catalogo/<slug>` para cualquier dataset registrado. Cada visor expone:
  - Controles de agregación dinámica (dimensión/métrica) y gráfico interactivo (Chart.js).
  - Tabla de muestra (primeras 100 filas) con enlace directo a la descarga `CSV`.
  - Trazabilidad (`lineaje`, inventario raw) y enlaces API permanentes.

9 Gendarmería (Raw/Bronze)

- `data/meta/catalog/gendarmeria_reportes_mensuales.yaml`: catálogo para los reportes mensuales de Gendarmería (subsistemas Cerrado, Abierto y Postpenitenciario), con cobertura 2019-2025 y notas sobre meses faltantes (sept-2023 y mar/abr-2024 según subsistema).
- Manifest: `data/meta/gendarmeria_reportes_mensuales_files_metadata.csv` (registro completo de descargas, hash SHA-256, tamaño y URL original). QA mínimo: verificar 404 detectados en la página de origen antes de marcar pendientes.
- Diccionarios:

  - data/meta/diccionarios/gendarmeria_unidades.csv (175 códigos de establecimiento con región asignada).
  - data/meta/diccionarios/gendarmeria_unidades_alias.csv (cambios de nomenclatura relevantes, por ejemplo CPF SAN MIGUEL vs CPF Mayor Marisol Estay).
  - data/meta/diccionarios/gendarmeria_abierto_medidas.csv, gendarmeria_cerrado_pobl_reclu gendarmeria_postpenitenciario_medidas.csv para codificar medidas por vista.
- Loader: etl/pipelines/gendarmeria_reportes/loader.py::GendarmeriaReportLoader genera tres salidas en data/bronze/gendarmeria/ (gendarmeria_abierto.csv, gendarmeria_cerrado_pob gendarmeria_postpenitenciario.csv). Cada dataset está normalizado (formato long), con nivel de registro (unidad, total_region, total_nacional) y metadatos de dimensión añadidos.
- Esquema propuesto en data/meta/schema/gendarmeria_reportes_mensuales.sql; pendiente desarrollar modelos silver/gold (dimensiones de unidades, regiones y tipologías delictuales).
- Consideraciones QA:
  - El HTML 2024 repite el vínculo de mayo para S.Postpenitenciario jun24; la descarga se omite hasta que la fuente publique el archivo correcto.
  - En S.Cerrado, las columnas con X indican presencia de unidad por régimen; se registran como tipo_valor=flag para diferenciarlas de conteos numéricos.
  - Revisar periódicamente los enlaces 404 (s.abierto_abr24.xlsx, s.postpenitenciario_mar24.xlsx, etc.) por si la fuente restituye los archivos.

10 ENUSC (Raw)

- data/meta/catalog/enusc.yaml: catálogo para la Encuesta Nacional Urbana de Seguridad Ciudadana (bases .sav, cuestionarios, manuales e insumos metodológicos).
- Manifest: data/meta/enusc_files_metadata.csv (hash, tamaño y categoría por recurso).
- Cobertura disponible: bases de usuario 2009-2024 (ediciones VI-XXI) y consolidados interanuales 2008-2021 / 2008-2022 / 2008-2024. La base 2008 (edición V) sigue ausente en la fuente pública.
- Repositorio raw incluye manuales de usuario, manuales de trabajo de campo, diccionarios de variables y tabulados 2016-2024 (data/raw/enusc/manuales*, diccionarios/, tabulados/, notas/).
- Pendiente definir pipeline bronze con variables homologadas y diccionarios de código (territorial, victimización, factores de riesgo).

11 Fiscalía (Raw → Bronze)

- data/meta/catalog/fiscalia_persecucion_penal.yaml: catálogo para los boletines institucionales, anuarios estadísticos y reportes temáticos publicados por la Fiscalía de Chile sobre persecución penal.
- Manifest: data/meta/fiscalia_persecucion_penal_files_metadata.csv (103 archivos descargados el 13-oct-2025: 70 PDF, 32 XLS y 1 XLSX con coberturas 2004-2025).

- Descargas origen: https://www.fiscaliadechile.cl/persecucion-penal/estadisticas (páginas 0-7 rastreadas con script ad-hoc).
- Bronze disponible:
- `data/bronze/fiscalia_persecucion_penal/fiscalia_metrics.csv`: tabla larga con todas las hojas `TB*`, medidas (`measure_label`, `measure_slug`) y dimensiones territoriales (`region_name`, `fiscalia_rm`).
- `data/bronze/fiscalia_persecucion_penal/fiscalia_annual_totals.csv`: totales nacionales (TOTAL / TOTAL NACIONAL) seleccionando el corte más reciente de cada año.
- `data/bronze/fiscalia_persecucion_penal/fiscalia_region_annual.csv`: series anuales por región (sin duplicar trimestres acumulados).
- `data/bronze/fiscalia_persecucion_penal/fiscalia_rm_annual.csv`: agregados anuales para las cuatro fiscalías de la RM (Centro Norte, Oriente, Occidente, Sur).
- Dimensiones territoriales interoperables con `data/geo_region_provincia_comuna.csv` (`region_code`, `region_iso`) y con el catálogo de categorías de delito (`data/meta/diccionarios/fiscalia_c`
- QA automático (`data/meta/fiscalia_persecucion_penal_quality_checks.csv`): controles de sumas regionales, proporciones y subtotales (sin advertencias tras aplicar tolerancias y excluir bloques porcentuales).

12 Referencias BCCh (IPC / PIB)

- `data/meta/catalog/ref_bcch.yaml`: metadatos del paquete de series provenientes del Banco Central de Chile.
- `scripts/fetch_bcch_series.py`: script que consume el servicio SOAP `GetSeries` (SieteWS) usando `BCCH_USER` y `BCCH_PASS`.
- Archivos generados:
  - `data/ref/bcch/ipc_mensual.csv`: IPC general histórico (base 2018=100, frecuencia mensual).
  - `data/ref/bcch/pib_trimestral_real_2018.csv`: PIB real encadenado (base 2018=100, CLP de 2018, frecuencia trimestral).
- Uso esperado:
  - Deflactar montos nominales antes de calcular variaciones reales.
  - Obtener porcentajes respecto del PIB para series agregadas (ej. presupuesto / PIB).
- Referencias y términos: https://si3.bcentral.cl/estadisticas/Principal1/Web_Services/index_BDE_TC.htm