



Informe Técnico: Revisión de Dominio de Datos



Contenido

1. Objetivos.....	3
Objetivo general	3
Objetivos específicos.....	3
2. Metodología.....	4
2.1. Proceso Metodológico	4
2.1.1 Extracción de nombres de campos asociados a dominios	4
2.1.2 Obtención de valores únicos usados por cada campo	5
2.1.3 Reconstrucción de los dominios definidos en la GDB	7
2.1.4 Persistencia de resultados como vistas materializadas	8
2.2. Componentes Desarrollados	9
2.2.1 Función <code>distinct_por_columna()</code>	9
2.2.2 Vista materializada estructura.....	10
2.2.3 Vista materializada estructura_data	11
CONCLUSIÓN.....	12

Índice de tablas

Tabla 1: Resultado de la función <code>distinct_por_columna</code>	5
Tabla 2: Resultado de la función estructura	6
Tabla 3:Reconstrucción de los dominios definidos en la GDB.....	8
Tabla 4: Script <code>distinct_por_columna</code>	9

Índice de Ilustraciones

Ilustración 1:Filtro por tipo de objeto y nombre.....	4
Ilustración 2:Generación de SQL dinámico	6
Ilustración 3:Resultados de la función <code>distinct_por_columna()</code> (valores únicos por campo con dominio).	7
Ilustración 4: Reusltados estructura_data	9
Ilustración 5: Script función estructura	11
Ilustración 6: Script función estructura_data.....	11

1. Objetivos

Objetivo general

Desarrollar un conjunto de funciones y vistas materializadas que permitan identificar, estructurar y analizar los valores únicos utilizados en campos con dominios en la base de datos, facilitando la validación y control de calidad de los datos geográficos integrados desde una Geodatabase de ArcGIS.

Objetivos específicos

- ✓ **Identificar dinámicamente** los campos con dominios declarados en las tablas y clases de entidad almacenadas en la Geodatabase de ArcGIS.
- ✓ **Extraer los valores únicos** almacenados en columnas asociadas a dominios, para conocer el uso real de los datos en producción.
- ✓ **Reconstruir la estructura completa** de los dominios definidos en la GDB, incluyendo los valores válidos asociados a cada columna.
- ✓ **Almacenar de forma persistente** tanto la estructura esperada como los datos utilizados mediante vistas materializadas para facilitar consultas posteriores.
- ✓ **Permitir comparaciones automáticas** entre los valores usados y los definidos en los dominios para detectar inconsistencias, errores de digitación u homologaciones pendientes.
- ✓ **Facilitar la trazabilidad** entre la implementación de modelos de datos espaciales y el uso efectivo de las reglas de negocio asociadas a los dominios en el sistema de gestión de base de datos.

2. Metodología

2.1. Proceso Metodológico

2.1.1 Extracción de nombres de campos asociados a dominios

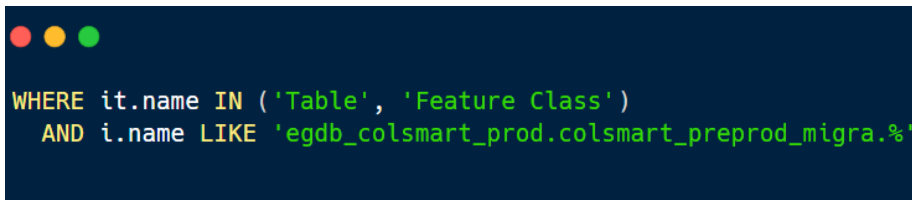
El primer paso del proceso metodológico consiste en identificar todos los campos que, dentro de las tablas y clases de entidad de la Geodatabase, tienen asignado un dominio. Esta información es fundamental para establecer una correspondencia entre los valores usados en los datos y las restricciones definidas en el modelo conceptual.

a. Fuente de datos: sde.gdb_items

ArcGIS almacena las definiciones estructurales de tablas, clases de entidad y dominios en la tabla sde.gdb_items, utilizando el campo definition, el cual contiene un XML con los metadatos del dataset. Este XML describe, entre otros aspectos, los nombres de los campos, sus tipos y los dominios asociados.

b. Filtro por tipo de objeto y nombre

Se filtran únicamente los objetos cuyo tipo corresponde a Table o Feature Class, ya que son los únicos que contienen campos editables con posibles dominios. Adicionalmente, se restringe la búsqueda a objetos cuyo nombre incluya el patrón colsmart_preprod_migra, correspondiente al esquema de trabajo.



```
WHERE it.name IN ('Table', 'Feature Class')
AND i.name LIKE 'egdb_colsmart_prod.colsmart_preprod_migra.%'
```

Ilustración 1: Filtro por tipo de objeto y nombre

c. Navegación en el XML

Mediante la función xpath() se accede a los nodos <GPFieldInfoEx> dentro del XML, que representan cada campo definido en la tabla o clase de entidad. Para cada uno de estos nodos, se extraen:

- **Nombre del campo:**
xpath('///Name/text()', cv_node)[1]::text
- **Nombre del dominio asignado:**
xpath('///DomainName/text()', cv_node)[1]::text

d. Normalización y partición del nombre completo

Los nombres de tabla recuperados desde sde.gdb_items.name están en formato egdb_colsmart_prod.colsmart_preprod_migra.nombre_tabla. Se aplica split_part() para separar:

- El **nombre del esquema**: split_part(..., '.', 2)

- El **nombre de la tabla**: `split_part(..., ',', 3)`

Todo se transforma a minúsculas para garantizar consistencia en las comparaciones posteriores.

e. Resultado del proceso

Como salida de esta etapa, se obtiene una estructura con las siguientes columnas:

Tabla 1: Resultado de la función `distinct_por_columna`

Campo	Descripción
<code>schema_name</code>	Nombre del esquema (ej. <code>colsmart_preprod_migra</code>)
<code>table_name</code>	Nombre de la tabla o clase de entidad
<code>column_name</code>	Campo que tiene un dominio asociado
<code>domainname</code>	Nombre del dominio asignado según el XML

Este conjunto de resultados es esencial para continuar con la extracción de los valores únicos realmente utilizados en la base de datos (ver sección 2.1.2), así como para cruzarlos con los valores válidos definidos en los dominios (ver sección 2.1.3).

2.1.2 Obtención de valores únicos usados por cada campo

Una vez identificados los campos que tienen dominios asignados en la base de datos, el siguiente paso consiste en extraer los valores únicos realmente utilizados en esos campos dentro de las tablas del esquema `colsmart_preprod_migra`. Esto permite contrastar la implementación práctica de los datos con las restricciones definidas por los dominios.

a. Enfoque dinámico

Dado que los nombres de las tablas y columnas varían, y que la consulta debe adaptarse a cada combinación identificada en la etapa anterior, se implementa una función en PostgreSQL con SQL dinámico:

`colsmart_preprod_migra.distinct_por_columna()`

Esta función recorre dinámicamente todas las combinaciones tabla-columna obtenidas previamente y genera sentencias SQL personalizadas para cada caso.

b. Generación de SQL dinámico

Dentro del cuerpo de la función, se utiliza `format()` para construir consultas seguras que obtienen los valores únicos de cada columna con dominio. Se utiliza `GROUP BY` en lugar de `DISTINCT` para mayor control y claridad:

```
dyn_sql := format(
  'SELECT %L, %L, %L, %I::text
   FROM   %I.%I
   GROUP BY %I',
  rec.schema_name,
  rec.table_name,
  rec.column_name,
  rec.column_name,
  rec.schema_name,
  rec.table_name,
  rec.column_name
);
```

Ilustración 2: Generación de SQL dinámico

Esto asegura que todos los valores distintos utilizados en la columna sean devueltos como texto, sin importar su tipo original (entero, cadena, booleano, etc.).

c. Estructura del resultado

La función retorna una tabla con la siguiente estructura:

Tabla 2: Resultado de la función estructura

Campo	Descripción
schema_name	Nombre del esquema
table_name	Nombre de la tabla
column_name	Nombre del campo asociado al dominio
valor	Valor distinto encontrado en esa columna

d. Persistencia para comparación

Los resultados de esta función se utilizan para construir una vista materializada llamada `estructura_data`, la cual almacena los valores únicos utilizados en la base de datos para cada campo con dominio. Esta vista se utiliza posteriormente para realizar validaciones cruzadas (ver sección 2.3.1).

A-Z schema_name	A-Z table_name	A-Z column_name	A-Z name_value
colsmart_preprod_migra	ilc_caracteristicasunidadconstruccion	tipo_tipologia	5021132_Institucional.Tipo_7
colsmart_preprod_migra	ilc_interesado	porcentaje_propiedad	77
colsmart_preprod_migra	ilc_fuenteadministrativa	tipo	SENTENCIA
colsmart_preprod_migra	ilc_interesado	porcentaje_propiedad	68
colsmart_preprod_migra	ilc_caracteristicasunidadconstruccion	conservacion_tipologia	Malo_4
colsmart_preprod_migra	ilc_interesado	porcentaje_propiedad	10
colsmart_preprod_migra	ilc_interesado	porcentaje_propiedad	73
colsmart_preprod_migra	ilc_caracteristicasunidadconstruccion	tipo_tipologia	1021125_Residencial.Tipo_5
colsmart_preprod_migra	ilc_predio	destinacion_economica	Infraestructura_Seguridad
colsmart_preprod_migra	ilc_caracteristicasunidadconstruccion	uso	Comercial_Teatro_Cinemas
colsmart_preprod_migra	ilc_interesado	porcentaje_propiedad	63

Ilustración 3: Resultados de la función `distinct_por_columna()` (valores únicos por campo con dominio).

2.1.3 Reconstrucción de los dominios definidos en la GDB

Una vez identificados los campos asociados a dominios y los valores únicos utilizados en los datos, es necesario reconstruir la estructura completa de los dominios definidos en la Geodatabase (GDB) de ArcGIS. Esta información permite establecer una base de comparación entre lo que está definido como válido y lo que realmente se ha utilizado en la base de datos.

ArcGIS almacena los dominios en la tabla `sde.gdb_items`, dentro del campo `definition`, que contiene un documento XML. En este XML se especifican los nombres de los dominios, así como los valores permitidos (coded values) para cada uno. A través de consultas XPath se accede a estos nodos, extrayendo tanto el nombre del dominio como sus valores definidos.

De cada nodo se obtiene el nombre del dominio con `xpath('//DomainName/text()', definition)` y se extraen los valores válidos con `xpath('//CodedValues/CodedValue', definition::xml)`, accediendo a los códigos mediante `xpath('//Code/text()', cv_node)`.

Estos valores se cruzan posteriormente con los nombres de campos obtenidos en la etapa 2.1.1, de forma que es posible saber exactamente qué columnas tienen dominios asignados y cuáles son los valores que deberían ser válidos en cada una de ellas.

Para almacenar esta información, se construye la vista materializada `colsmart_preprod_migra.estructura`, que representa la estructura lógica completa de dominios: columnas con dominio y sus valores válidos. Esta vista combina los valores extraídos de los dominios (dominios) con los campos identificados previamente (conecta), mediante una unión interna por el nombre del dominio.

Esta vista sirve como referencia oficial de lo que se espera en cada campo controlado por un dominio, y será utilizada posteriormente en los procesos de validación (sección 2.3) para identificar discrepancias y errores en los datos.

Tabla 3:Reconstrucción de los dominios definidos en la GDB

A? schema_name	A? table_name	A? column_name	A? name_value
colsmart_preprod_migra	cr_fuenteespacial	tipo_fuente_espacial	Croquis_Campo
colsmart_preprod_migra	cr_fuenteespacial	tipo_fuente_espacial	Datos_Crudos
colsmart_preprod_migra	cr_fuenteespacial	tipo_fuente_espacial	Informe_Tecnico
colsmart_preprod_migra	cr_fuenteespacial	tipo_fuente_espacial	Ortofoto
colsmart_preprod_migra	cr_fuenteespacial	tipo_fuente_espacial	Registro_Fotografico
colsmart_preprod_migra	ilc_interesado	tipo	Persona_Juridica
colsmart_preprod_migra	ilc_interesado	tipo	Persona_Natural

2.1.4 Persistencia de resultados como vistas materializadas

Con el objetivo de optimizar las consultas posteriores y permitir validaciones eficientes, los resultados de los procesos anteriores se almacenan en dos vistas materializadas: estructura y estructura_data. Estas vistas representan, respectivamente, los valores **definidos** y los **utilizados** en campos controlados por dominios.

La persistencia en vistas materializadas garantiza que:

- La información está disponible de forma inmediata sin necesidad de ejecutar funciones complejas o recorridos XML cada vez.
- Se pueden realizar comparaciones cruzadas de forma ágil, incluso sobre volúmenes grandes de datos.
- Se puede programar la actualización (REFRESH) de forma periódica o desencadenada por cambios en los datos.

La vista colsmart_preprod_migra.estructura contiene todos los valores válidos definidos para cada campo con dominio, obtenidos desde el XML de la Geodatabase. Por su parte, la vista colsmart_preprod_migra.estructura_data contiene los valores únicos que han sido realmente utilizados en la base de datos, recuperados mediante la función dinámica `distinct_por_columna()`.

Esta simetría entre ambas estructuras permite construir consultas de validación para identificar:

- Valores usados que no están definidos en el dominio (errores).
- Dominios definidos que no han sido utilizados (potencial limpieza o revisión).
- Inconsistencias entre el diseño de la GDB y los datos efectivamente registrados.

La estrategia de persistencia mediante vistas materializadas también facilita la integración de este módulo dentro de flujos ETL automatizados o tareas de validación periódica, ya sea desde procesos manuales, reportes de calidad, o validadores automáticos como Great Expectations u otros sistemas de control.

Az schema_name	Az table_name	Az column_name	Az name_value
colsmart_preprod_migra	ilc_caracteristicasunidadconstruccion	tipo_tipologia	5021132_Institucional.Tipo_7
colsmart_preprod_migra	ilc_interesado	porcentaje_propiedad	77
colsmart_preprod_migra	ilc_fuenteadministrativa	tipo	SENTENCIA
colsmart_preprod_migra	ilc_interesado	porcentaje_propiedad	68
colsmart_preprod_migra	ilc_caracteristicasunidadconstruccion	conservacion_tipologia	Malo_4
colsmart_preprod_migra	ilc_interesado	porcentaje_propiedad	10
colsmart_preprod_migra	ilc_interesado	porcentaje_propiedad	73
colsmart_preprod_migra	ilc_caracteristicasunidadconstruccion	tipo_tipologia	1021125_Residencial.Tipo_5
colsmart_preprod_migra	ilc_predio	destinacion_economica	Infraestructura_Seguridad
colsmart_preprod_migra	ilc_caracteristicasunidadconstruccion	uso	Comercial_Teatro_Cinemas

Ilustración 4: Reusitados estructura_data

2.2. Componentes Desarrollados

2.2.1 Función distinct_por_columna().

Esta función recorre dinámicamente todas las tablas del esquema **colsmart_preprod_migra** que poseen un dominio declarado en la Geodatabase de ArcGIS y devuelve los **valores distintos realmente utilizados** en cada columna controlada por dominio.

El resultado sirve como “foto” de la realidad de los datos y se materializa en la vista estructura_data para acelerar validaciones posteriores.

Tabla 4: Script distinct_por_columna

Nº	Bloque	¿Qué hace?	¿Por qué es necesario?
1	“Recoger columnas con dominio”	Lee el XML de ArcGIS (sde.gdb_items) para encontrar todos los campos que declaran un DomainName. Por cada campo obtiene esquema, tabla y nombre de la columna.	<pre> CREATE OR REPLACE FUNCTION colsmart_preprod_migra.distinct_por_columna() RETURNS TABLE (schema_name text, table_name text, column_name text, valor text) LANGUAGE plpgsql AS \$\$ DECLARE rec record; dyn_sql text; BEGIN /* --- 1. Recorre todos los campos con dominio que aparecen en el XML --- */ FOR rec IN SELECT lower(split_part(dataset_name, '.', 2)) AS schema_name, lower(split_part(dataset_name, '.', 3)) AS table_name, lower(xpath('//Name/text()', cv_node))[1]::text AS column_name FROM (SELECT i.name AS dataset_name, unnest(xpath('//GPFieldInfoExs/GPFieldInfoEx', i.definition::xml)) AS cv_node FROM sde.gdb_items i JOIN sde.gdb_itemtypes it ON it.uuid = i.type WHERE it.name IN ('Table', 'Feature Class') AND i.name LIKE 'egdb_colsmart_prod.colsmart_preprod_migra.%') campos_con_dominios WHERE (xpath('//DomainName/text()', cv_node))[1] IS NOT NULL LOOP </pre>
2	“Comprobar que la columna existe”	Consulta el catálogo information_schema.columns. Si la columna fue borrada o renombrada en la tabla (pero sigue apareciendo en el XML), se detecta la	<pre> /* --- 2. COMPROBAR QUE LA COLUMNA EXISTE --- */ PERFORM 1 FROM information_schema.columns c WHERE c.table_schema = rec.schema_name AND c.table_name = rec.table_name AND c.column_name = rec.column_name; IF NOT FOUND THEN CONTINUE; -- si el XML habla de una columna que no existe, la ignoramos END IF; </pre>

```

/* ----- 3. EXTRAER VALORES
ÚNICOS ----- */
    dyn_sql := format(
        'SELECT %L, %L, %L, %I::text
        FROM %I.%I
        GROUP BY %I',
        rec.schema_name,           -- se devuelven
        como literales
        rec.table_name,
        rec.column_name,
        rec.column_name,         -- identificador
        a castear a texto
        rec.schema_name,
        rec.table_name,
        rec.column_name
    );

    RETURN QUERY EXECUTE dyn_sql; -- añade el
    resultado al conjunto que devolverá la función
END LOOP;

END;
$$;
```

```

CREATE MATERIALIZED VIEW colsmart_preprod_migra.estructura AS
WITH dominios AS (
    -- 1 Extraer códigos permitidos
    SELECT
        domainname,
        (xpath('//Code/text()', cv_node))[1]::text AS name_value
    FROM (
        SELECT
            (xpath('//DomainName/text()', definition))[1]::text AS domainname,
            unnest(xpath('//CodedValues/CodedValue', definition::xml)) AS cv_node
        FROM sde.gdb_items
        WHERE (xpath('//Owner/text()', definition))[1]::text = 'colsmart_preprod_migra'
    ) coded_values_nodes
),
conecta AS (
    -- 2 Localizar qué tabla-columna usa cada dominio
    SELECT
        lower(split_part(dataset_name, '.', 2)) AS schema_name,
        lower(split_part(dataset_name, '.', 3)) AS table_name,
        (xpath('//Name/text()', cv_node))[1]::text AS column_name,
        (xpath('//DomainName/text()', cv_node))[1]::text AS domainname
    FROM (
        SELECT i.name AS dataset_name,
            unnest(xpath('//GPFieldInfoExs/GPFieldInfoEx',
                i.definition::xml)) AS cv_node
        FROM sde.gdb_items i
        JOIN sde.gdb_itemtypes it ON it.uuid = i.type
        WHERE it.name IN ('Table', 'Feature Class')
            AND i.name LIKE 'egdb_colsmart_prod.colsmart_preprod_migra.%'
    ) coded_values_nodes
    WHERE (xpath('//DomainName/text()', cv_node))[1] IS NOT NULL
)
SELECT c.schema_name,
    c.table_name,
    c.column_name,
    d.name_value
FROM conecta c
JOIN dominios d
ON d.domainname = c.domainname;

```

Ilustración 5: Script función estructura

2.2.3 Vista materializada estructura_data

La vista materializada estructura convierte los dominios de ArcGIS en una tabla de valores permitidos por columna. Sirve como referencia normativa para las validaciones de datos. El script que la crea se muestra a continuación.

```

CREATE MATERIALIZED VIEW colsmart_preprod_migra.estructura_data AS
select distinct c.schema_name,c.table_name,c.column_name,c.valor as name_value
FROM colsmart_preprod_migra.distinct_por_columna() c;

```

Ilustración 6: Script función estructura_data

CONCLUSIÓN

La solución propuesta automatiza la reconciliación entre los dominios definidos en la geodatabase ArcGIS y los valores realmente almacenados en PostgreSQL: la función `distinct_por_columna()` inspecciona los metadatos XML, valida la existencia física de cada columna y genera dinámicamente las consultas que extraen los valores distintos; las vistas materializadas `estructura` y `estructura_data` persisten, respectivamente, la definición normativa de cada dominio y la “fotografía” operativa de los datos, de modo que cualquier auditoría, validación ETL o análisis de consistencia pueda ejecutarse con rapidez y sin recargar el servidor; en conjunto, el diseño modular hace escalable y mantenible el control de calidad al ofrecer un puente robusto, reproducible y fácil de integrar entre el modelo conceptual y la práctica diaria de los datos espaciales.