ANH

AGENCIA NACIONAL DE HIDROCARBUROS

Contratar el análisis, diseño, implementación y puesta en funcionamiento del Banco de Información Energética (BIEN(Banco de Información Energética)) que contemple la gobernanza del dato y permita la carga de información, la estandarización, el acceso y el manejo integral y sistemático de la información gestionada por la Vicepresidencia Técnica de la ANH(Agencia Nacional de Hidrocarburos).



Informe Desarrollo Sistema Gestión Datos

Versión 1.0

Periodo del informe:

13 a 19 de Enero de 2026

Fecha de presentación:

21 de Enero de 2026

Bogotá, Colombia

DOCUMENTO CONFIDENCIAL

UT ANH PWD – NX 2025

Unión Temporal ANH PWD-NX 2025

Historial de versiones

Id	Fecha (dd/mm/aa)	Versión	Descripción	Elaborado por	Revisado por
1	22/01/2026	1.0	Versión inicial del informe	UT ANH PWD – NX 2025	Gerente del Proyecto

CONFIDENCIAL

Unión Temporal ANH PWD-NX 2025

Contenido

1. Introducción 6
2. Objetivos del Documento 6
2.1 Objetivo General 6
2.2 Objetivos Específicos 6
3. Alcance del Informe 7
3.1 Incluye 7
3.2 Excluye 7
3.3 Supuestos y restricciones 7
4. Marco de Referencia del Desarrollo 8
4.1 Trazabilidad documental y línea base 8
4.2 Metodología de construcción y aseguramiento 9
4.3 Principios y decisiones de arquitectura aplicadas 9
5. Descripción General de la Plataforma Tecnológica 9
5.1 Tipos de información y volumen de referencia 10
5.2 Perfiles de usuario y responsabilidades 10
5.3 Capacidades funcionales implementadas (visión por requerimientos) 10
5.4 Componentes principales de la solución 11
6. Arquitectura del Sistema de Gestión de Datos 11
6.1 Vista general de la arquitectura de datos del BIEN 12
6.2 Estrategia de persistencia y bases de datos 13
6.2.1 Bases de datos para información estructurada 13
6.2.2 Repositorios para información no estructurada 14
6.2.3 Almacenamiento de información geoespacial 14
6.3 Capa de indexación y motores de búsqueda del BIEN 14
6.3.1 Búsqueda sobre datos estructurados 15
6.3.2 Búsqueda sobre información no estructurada (full-text) 15
6.3.3 Indexación semántica y búsqueda avanzada 16

CONFIDENCIAL

6.4 Relación entre repositorios, catálogo e índices de búsqueda 17
6.5 Componentes y responsabilidades (nivel contenedores): 17
7. Desarrollo de Componentes de la Plataforma 19
7.1 Ingesta y pipelines (ETL/ELT) 19
7.2 Ingesta y pipelines (ETL/ELT) – detalle ampliado 19
7.3 Repositorio unificado y modelo de datos 21
7.4 Catálogo, metadatos, taxonomía y linaje 22
7.5 Búsqueda e indexación (full-text / vectorial) y NLQ/RAG 22
7.6 GIS: geovisor y servicios geoespaciales 23
7.7 Calidad, pruebas y desempeño 23
8. Gestión de Datos y Gobernanza 24
8.1 Metadatos mínimos y clasificación 24
8.2 Linaje y retención 24
9. GESTION DE DATOS Y DECISIONES DE DISEÑO 24
9.1 ¿Por qué SQL Server? 25
9.2. ¿Por qué MongoDB? 26
9.3 Sincronización entre Plataformas 26
9.4 . Decisiones arquitectónicas 27
10. Seguridad, Control y Trazabilidad 35
10.1 Auditoría y registros 35
10.2 Observabilidad 36
10.3 DevSecOps 36
11. Integraciones e Interoperabilidad 36
11.1 Workflow de integración y control 37
11.2 Contratos técnicos 37
11.3 Integración con herramientas especializadas 38
11.4 Catálogo de servicios e integración (resumen operativo) 38
12. Estrategia de Implementación y Despliegue 39

CONFIDENCIAL

12.1 Fase I — Preparación & Ready-to-Build 39
12.2 Fase II — Construcción y Despliegues Iterativos 39
12.3 Fase III — Pruebas Integrales, Capacitación y Corte a Producción 39
12.4 Gestión por oleadas (MVP e incrementos) 40
12.5 Controles de liberación (Go/No-Go) y rollback 40
13. Consideraciones de Operación y Evolución 40
13.1 Escalabilidad y crecimiento 41
13.2 Operación temprana 41
13.3 Operación, soporte y continuidad 41
13.4 Evolución y mejoras 41
14. Conclusiones 42
15. Anexos 42
15.1 Anexo 1 - Infraestructura de datos estructurados no estructurados 42
15.2 Anexo 2 – Diccionario de datos 42
15.3 Anexo 3 - Diagrama MER SQLServer 42
15.4 Anexo 4 - Diagrama MongoDB 42
16. Glosario 42

CONFIDENCIAL

Unión Temporal ANH PWD-NX 2025

1. Introducción

El presente Informe de Desarrollo de la Plataforma Tecnológica se elabora en cumplimiento de los pliegos del proceso ANH-002-CM-2025 y del alcance contractual asociado al Banco de Información Energética (BIEN). El documento consolida, de manera técnica y auditable, cómo se diseñó, construyó, integró y dispuso la solución tecnológica que habilita el Sistema de Gestión de Datos del BIEN.

El BIEN se concibe como una plataforma institucional para centralizar, organizar y gobernar la información técnica, documental y geoespacial de la Vicepresidencia Técnica de la ANH, permitiendo su consulta, análisis, trazabilidad y explotación segura, con reglas de clasificación y responsables claramente definidos.

Este informe no introduce requerimientos nuevos. Su propósito es describir el desarrollo e implementación de la plataforma, evidenciando su alineación con los requerimientos funcionales, técnicos y no funcionales, así como con las decisiones arquitectónicas aprobadas durante la ejecución del proyecto.

2. Objetivos del Documento

2.1 Objetivo General

Documentar el desarrollo de la plataforma tecnológica del BIEN, describiendo la arquitectura implementada, los componentes construidos, los mecanismos de gestión y gobierno de datos, y los controles de seguridad, trazabilidad e interoperabilidad requeridos por la ANH.

2.2 Objetivos Específicos

- Describir la arquitectura de referencia y su materialización en la solución implementada.

- Detallar los componentes del Sistema de Gestión de Datos: ingesta, repositorio, catálogo, búsqueda, GIS, analítica e IA, y su integración.

CONFIDENCIAL

- Evidenciar los mecanismos de control: seguridad, auditoría, inmutabilidad, observabilidad, calidad y continuidad (DRP).

- Presentar la estrategia de despliegue por ambientes y el enfoque iterativo de construcción y validación (Go/No-Go).

3. ALCANCE DEL INFORME

3.1 Incluye

- Desarrollo e implementación del Sistema de Gestión de Datos del BIEN (componentes, integraciones y controles).

-Configuración de ambientes (DEV, QA, PRE, PROD) y habilitación de automatización de despliegue.

- Implementación de mecanismos de gobernanza (metadatos, taxonomía, linaje y auditoría).

- Implementación de mecanismos de seguridad (RBAC/ABAC, MFA/SSO cuando aplique, cifrado y gestión de secretos).

- Estrategia de pruebas, aceptación y evidencias asociadas.

3.2 Excluye

- Aspectos administrativos, presupuestales o de evaluación contractual que no correspondan al desarrollo técnico del sistema.

- Definición de políticas institucionales no aprobadas (se documenta su aplicación dentro del sistema cuando existan).

- Desarrollo de herramientas transaccionales ajenas al alcance del BIEN.

3.3 Supuestos y restricciones

La solución se integra con la infraestructura tecnológica institucional según lineamientos de TI de la ANH y soporta el crecimiento de volumen de información

CONFIDENCIAL

y demanda de usuarios conforme a los requerimientos definidos. Las integraciones se activan únicamente con autorización formal y evidencia verificable de intercambio.

4. Marco de Referencia del Desarrollo

El desarrollo se fundamenta en los siguientes documentos y lineamientos del proyecto BIEN:

- Pliegos del Concurso de Méritos ANH-002-CM-2025 y anexos técnicos.

- Contrato y Anexo Técnico del BIEN (características técnicas generales del objeto).

- Informe de Levantamiento de Requerimientos (RF, RT, RNF y criterios de aceptación).

- Documentos de Requerimientos Funcionales, Técnicos y No Funcionales.

- Documento de Arquitectura de Software (SAD) y arquitectura preliminar del BIEN.

- Plan de Implementación y Estrategia de Despliegue (WBS/EDT, oleadas, ambientes y pruebas).

Adicionalmente, se consideran estándares y buenas prácticas aplicables: ISO/IEC 25010 (calidad de software), OGC (servicios geoespaciales), ISO 19115/19139 e ISO 11179 (metadatos), WCAG 2.1 AA (accesibilidad), y principios de gobernanza alineados a DAMA-DMBOK y COBIT.

4.1 Trazabilidad documental y línea base

El desarrollo se ejecutó con una línea base controlada de requerimientos y artefactos, asegurando trazabilidad REQ-Diseño-Construcción-Pruebas-Evidencia. Para ello se consolidó un inventario de fuentes, un catálogo inicial de

CONFIDENCIAL

dominios y un mapeo de requerimientos por capacidades (búsqueda, documental, GIS, analítica, IA, integraciones, identidad y auditoría).

Como insumos de referencia se consideraron, entre otros, el Informe de Levantamiento de Requerimientos, los documentos de Requerimientos Funcionales (RF-01 a RF-10), Requerimientos No Funcionales (RNF-01 a RNF-09) y Requerimientos Técnicos (RT-01 a RT-09), el Documento de Arquitectura (SAD) y el Informe de Arquitectura Preliminar, así como el Plan de Implementación y Estrategia de Despliegue y el catálogo de APIs/servicios.

4.2 Metodología de construcción y aseguramiento

Se adoptó un enfoque iterativo e incremental, apoyado en gestión de backlog y control de cambios, con validaciones por ambiente (DEV/QA/PRE/PROD) y revisiones técnicas cruzadas. Cada incremento incorporó criterios de aceptación verificables, evidencias de pruebas y actas de Go/No-Go para el avance a la siguiente oleada.

4.3 Principios y decisiones de arquitectura aplicadas

Las decisiones de arquitectura se alinearon con principios de modularidad, separación de responsabilidades, seguridad por diseño y gobernanza del dato. Se incorporaron vistas C4 (Contexto/Contenedores) y vistas 4+1 (lógica, desarrollo, procesos, despliegue y escenarios), junto con una organización de componentes tipo arquitectura hexagonal para facilitar la evolución, pruebas y mantenibilidad.

5. Descripción General de la Plataforma Tecnológica

La plataforma BIEN implementa un Sistema de Gestión de Datos para administrar información estructurada, no estructurada y geoespacial asociada a dominios G&G&E. El sistema soporta la ingesta controlada con validaciones, el almacenamiento en repositorio central, la publicación en catálogo de

CONFIDENCIAL

metadatos con taxonomía oficial, la búsqueda unificada e indexación (full-text y vectorial), la visualización geográfica mediante servicios GIS/OGC y el consumo analítico mediante tableros y reportes.

5.1 Tipos de información y volumen de referencia

La solución está preparada para gestionar, entre otros, datos geofísicos y geológicos, modelos, sensores remotos, productos GIS y documentación digital, soportando formatos industriales y formatos geoespaciales. El volumen histórico a incorporar se referencia en centenas de miles de archivos y un orden de magnitud de cientos de terabytes.

5.2 Perfiles de usuario y responsabilidades

La operación se soporta en perfiles tipo: Administrador (gobierno y permisos), Oficiales de carga (ingesta y control de calidad), Analistas/Intérpretes (consumo especializado y generación de productos derivados) y Usuarios de consulta (acceso controlado a información autorizada).

5.3 Capacidades funcionales implementadas (visión por requerimientos)

En la plataforma se materializan capacidades principales alineadas con el levantamiento y el conjunto RF:

RF-01 Búsqueda unificada y workspace temporal: consultas por texto, filtros, facetas y guardado de "carpetas de búsqueda" como evidencia.

RF-02 Gestión documental y versionamiento: registro de activos, control de versiones, retención y auditoría asociada.

RF-03 Geovisor GIS: consumo de capas y consultas espaciales, con servicios estándar para interoperabilidad.

CONFIDENCIAL

RF-04 Analítica y reportes: KPIs, capa semántica, tableros oficiales y exportables (PDF/CSV) por rol.

RF-05 Asistente IA on-prem con RAG: consultas en lenguaje natural con recuperación de evidencia y citación.

RF-06 Tolerancia a fallos y recuperación: mecanismos de respaldo, restauración y pruebas de continuidad.

RF-07 Gestión de integraciones: APIs/OGC y workflow de solicitud–aprobación–evidencia de intercambio.

RF-08 Identidad, acceso y auditoría: control por roles/atributos y registro auditable de operaciones críticas.

RF-09 Ingesta continua post-entrega: operación evolutiva con estándares de carga y control de calidad.

5.4 Componentes principales de la solución

La plataforma se compone de: (i) canales de interacción (portal de consulta y backoffice), (ii) capa de servicios (APIs y orquestación), (iii) ingesta y procesamiento (ETL/ELT y validaciones), (iv) repositorios (datos estructurados, no estructurados y geoespaciales), (v) catálogo/metadata/taxonomía/linaje, (vi) búsqueda e indexación híbrida (texto + semántica), (vii) GIS (geovisor y servicios OGC), (viii) analítica/BI (capa semántica, KPIs y tableros) y (ix) controles transversales de seguridad, auditoría y observabilidad.

6. Arquitectura del Sistema de Gestión de Datos

La arquitectura del Sistema de Gestión de Datos del BIEN se diseñó para soportar el ciclo de vida completo de la información institucional, desde su ingesta y estandarización hasta su consulta, análisis y explotación avanzada, garantizando principios de gobernanza, seguridad, escalabilidad y desacoplamiento funcional.

CONFIDENCIAL

El diseño parte del reconocimiento de que los distintos tipos de información gestionados por el BIEN (datos estructurados, información no estructurada y datos geoespaciales) presentan naturalezas, volúmenes y patrones de uso heterogéneos, lo que hace inviable un enfoque monolítico de persistencia y consulta. En respuesta, la arquitectura adopta un modelo de persistencia especializada, complementado por una capa de indexación y búsqueda desacoplada, que optimiza el acceso sin comprometer la integridad ni el desempeño de las fuentes primarias.

6.1 Vista general de la arquitectura de datos del BIEN

Desde una perspectiva lógica, la arquitectura de datos del BIEN se organiza en capas claramente diferenciadas, cada una con responsabilidades específicas y controles asociados:

Capa de ingesta y procesamiento : Responsable de la incorporación controlada de información, validaciones técnicas, control de calidad y estandarización inicial.

Capa de persistencia y repositorios : Almacena los datos conforme a su naturaleza (estructurada, no estructurada, geoespacial), aplicando políticas de versionamiento, retención e integridad.

Capa de catalogación y metadatos : Pública y gobierna los activos de información mediante taxonomía, linaje y clasificación institucional.

Capa de indexación y búsqueda Construye índices optimizados para recuperación eficiente, sin impactar las bases de datos primarias.

Capa de exposición y consumo Habilita consultas, visualización GIS, analítica e interacción mediante APIs y portales.

Este enfoque permite que cada capa evolucione de forma controlada e independiente, manteniendo coherencia con los principios de modularidad y separación de responsabilidades definidos en la arquitectura preliminar.

CONFIDENCIAL

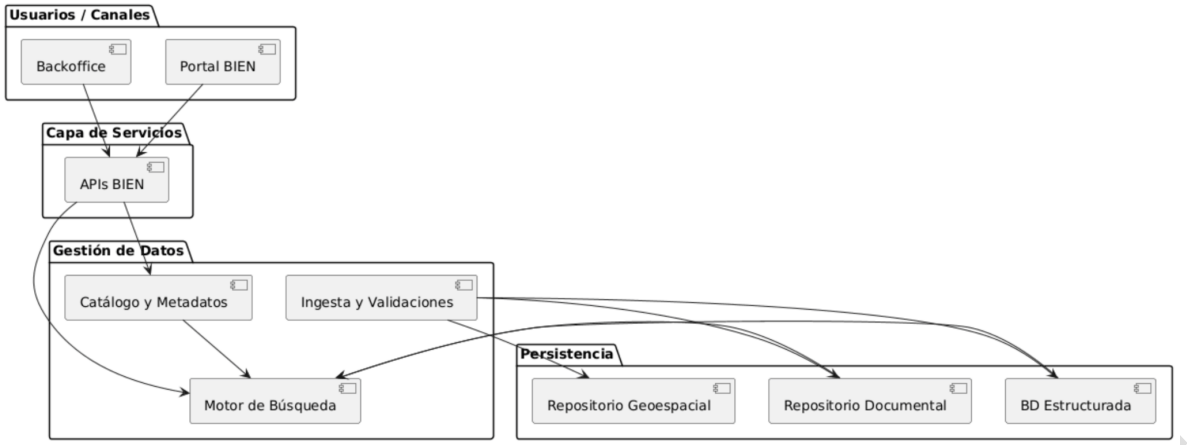


Diagrama 1. Vista general de arquitectura de datos

6.2 Estrategia de persistencia y bases de datos

La arquitectura del BIEN define una estrategia de persistencia distribuida, en la cual cada repositorio cumple un rol específico dentro del ecosistema de datos. Esta decisión responde a criterios de desempeño, escalabilidad, seguridad y gobernanza, y evita la sobrecarga de un único motor para todos los casos de uso.

6.2.1 Bases de datos para información estructurada

Los datos estructurados del BIEN se gestionan mediante motores de base de datos relacionales, responsables de garantizar:

- Integridad referencial y consistencia transaccional.

- Control de versiones lógicas cuando aplica.

- Soporte a modelos normalizados y semánticamente consistentes.

Estos motores no están diseñados para búsquedas exploratorias ni consultas intensivas por texto, razón por la cual su exposición a los usuarios finales se realiza exclusivamente a través de servicios y mecanismos de indexación derivados.

CONFIDENCIAL

6.2.2 Repositorios para información no estructurada

La información no estructurada (documentos técnicos, informes, archivos digitales) se almacena en repositorios documentales especializados, que permiten:

- Gestión de grandes volúmenes de archivos.

- Versionamiento y control de cambios.

- Asociación obligatoria de metadatos descriptivos y funcionales.

- Aplicación de políticas de clasificación y retención.

El contenido binario no se consulta directamente, sino que es procesado para extraer información relevante que alimente los índices de búsqueda.

6.2.3 Almacenamiento de información geoespacial

Los datos geoespaciales se gestionan en repositorios especializados que soportan:

- Sistemas de referencia espacial.

- Geometrías y atributos espaciales.

- Integración con servicios OGC y geovisores.

Este almacenamiento permite consultas espaciales especializadas, sin interferir con los repositorios documentales o relacionales.

6.3 Capa de indexación y motores de búsqueda del BIEN

La arquitectura del BIEN incorpora una capa de indexación desacoplada de las bases de datos, cuyo objetivo es optimizar la recuperación de información y habilitar capacidades avanzadas de búsqueda sin comprometer la estabilidad de las fuentes primarias.

CONFIDENCIAL

Esta capa consume información exclusivamente desde activos publicados y gobernados, y nunca desde datos en estado transaccional o no validados.

6.3.1 Búsqueda sobre datos estructurados

Para los datos estructurados, el motor de búsqueda opera sobre índices derivados, construidos a partir de:

- Metadatos funcionales.

- Campos relevantes previamente definidos.

- Relaciones semánticas asociadas a dominio, temporalidad y geografía.

El resultado de una búsqueda estructurada no es una consulta directa a tablas, sino la localización de datasets y vistas lógicas autorizadas, preservando el modelo de gobierno del BIEN.

6.3.2 Búsqueda sobre información no estructurada (full-text)

La búsqueda documental se apoya en mecanismos de indexación full-text, que permiten:

- Analizar el contenido textual de documentos.

- Asociar dicho contenido a metadatos y taxonomía institucional.

- Aplicar filtros y facetas por dominio, estado y clasificación.

Este enfoque facilita la localización eficiente de información técnica, manteniendo trazabilidad completa del activo original.

CONFIDENCIAL

6.3.3 Indexación semántica y búsqueda avanzada

Como complemento al full-text, la arquitectura contempla capacidades de indexación semántica, mediante representaciones vectoriales del contenido autorizado. Estas capacidades permiten:

- Mejorar la relevancia de los resultados.

- Soportar búsquedas por similitud conceptual.

- Habilitar consultas en lenguaje natural cuando aplique.

La indexación semántica no reemplaza los mecanismos tradicionales, sino que los potencia dentro de un marco controlado y auditable.

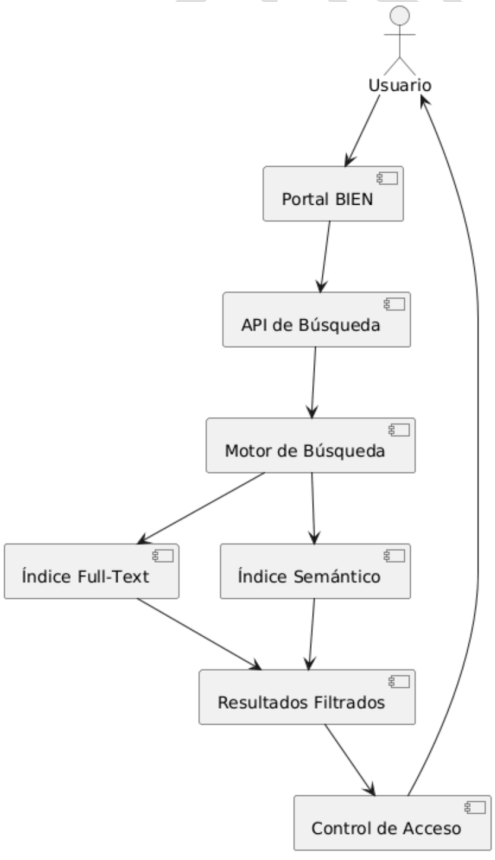


Diagrama 2. Flujo de indexación y búsqueda

6.4 Relación entre repositorios, catálogo e índices de búsqueda

El catálogo de metadatos actúa como eje articulador entre los repositorios físicos y los motores de búsqueda. Ningún activo es indexado ni expuesto sin cumplir previamente con:

- Registro completo de metadatos mínimos.

- Asignación de taxonomía institucional.

- Definición de responsable y clasificación.

- Estado formal de publicación.

De esta forma, la búsqueda en el BIEN se consolida como un mecanismo gobernado, en el que solo la información validada, autorizada y trazable es susceptible de ser recuperada y explotada.

6.5 Componentes y responsabilidades (nivel contenedores):

- Portal de consulta / experiencia de usuario: navegación por dominios, búsqueda, visualización documental y resultados GIS/analíticos.

- Backoffice: administración de cargas, metadatos, taxonomía, permisos, aprobaciones y publicación.

- Capa de servicios (APIs): endpoints versionados para búsqueda, catálogo, ingesta, auditoría, identidad y consumos especializados.

- Orquestación de ingesta (ETL/ELT): validación de estructura, control de calidad, normalización y registro de bitácoras.

- Repositorio de datos: almacenamiento para datos estructurados (motor relacional) y no estructurados (objetos/documentos), con versionamiento y retención.

- Servicios de catálogo y metadatos: publicación, taxonomía, linaje y trazabilidad extremo a extremo.

CONFIDENCIAL

- Motor de búsqueda: indexación textual (full-text) e indexación semántica (vectorial) para recuperación híbrida.

- GIS: servicios OGC (p.ej., WMS/WFS) y geovisor para capas temáticas y consultas espaciales.

- Analítica/BI: capa semántica (métricas/dimensiones) y tableros con control de acceso.

- IA (RAG/NLQ): componente de consulta en lenguaje natural con recuperación estricta y citación.

- Observabilidad y seguridad: registros, métricas y trazas; integración con SIEM; gestión de secretos; cifrado en tránsito y en reposo.

Nota: La arquitectura incorpora las vistas 4+1 y el enfoque de modularidad/hexagonal para separar puertos (APIs/ingesta/consumo) y adaptadores (repositorios, índices, GIS, autenticación), reduciendo el acoplamiento y facilitando pruebas y evolución.

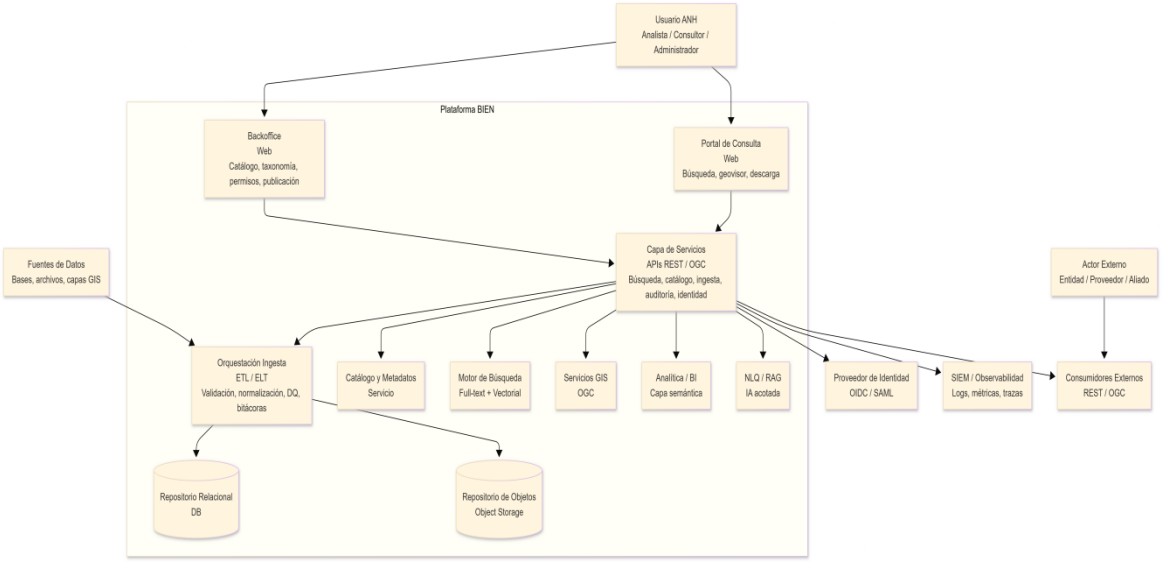


Diagrama 3. Diagrama de contenedores para búsqueda.

7. Desarrollo de Componentes de la Plataforma

7.1 Ingesta y pipelines (ETL/ELT)

Se implementan mecanismos de carga controlada para datos y documentos, incorporando validaciones de formato, control de calidad y clasificación. Se registran bitácoras de carga y eventos para garantizar trazabilidad técnica y funcional.

7.2 Ingesta y pipelines (ETL/ELT) – detalle ampliado

El componente de ingesta implementa un flujo controlado para incorporación de activos (datasets, documentos y productos GIS), garantizando consistencia técnica, metadatos mínimos y trazabilidad.

Flujo de ingesta (alto nivel):

1) Solicitud de carga y prevalidación: identificación del dominio, tipo de activo, responsable y clasificación.

2) Validaciones técnicas: estructura, formatos permitidos, tamaño, integridad (hash), CRS/proyección cuando aplique, y controles de duplicidad/versionamiento.

3) Validaciones de calidad: reglas por dominio (nulos/outliers, rangos, coherencia), y conformidad con estándares de carga/reporte.

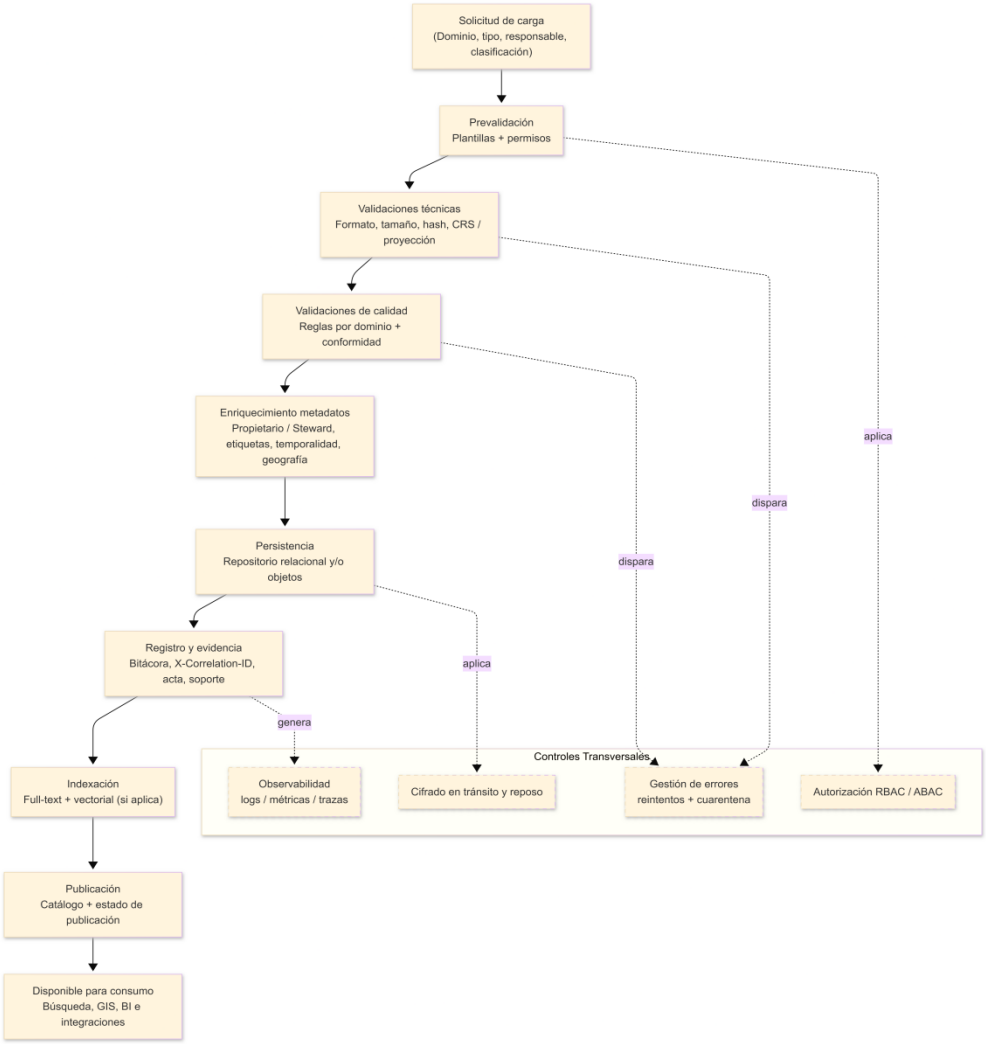
4) Enriquecimiento de metadatos: propietario, steward, descripción, etiquetas, temporalidad, geografía, sensibilidad y relaciones (proyecto/pozo/contrato, etc.).

5) Persistencia y registro: escritura en repositorios correspondientes, registro de bitácora y generación de eventos de auditoría.

6) Publicación: registro en catálogo, indexación (texto y/o vectores), y disponibilidad para búsqueda/consumo.

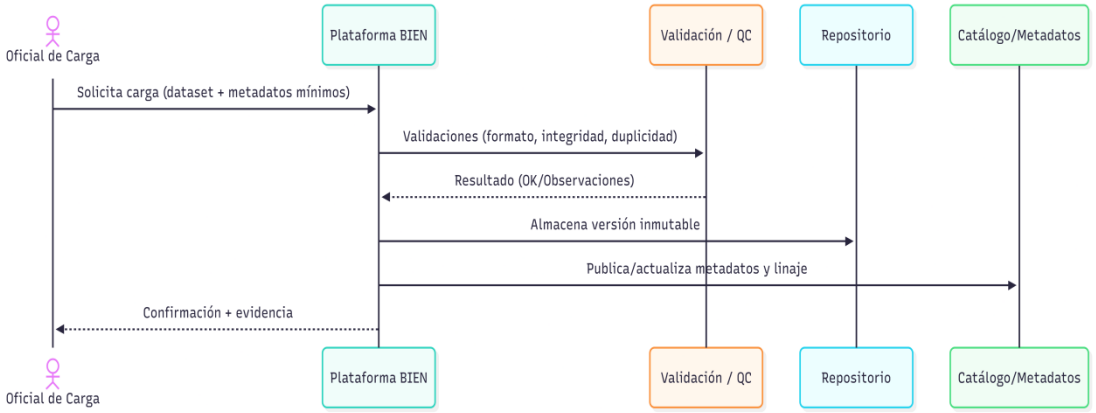
CONFIDENCIAL

Unión Temporal ANH PWD-NX 2025



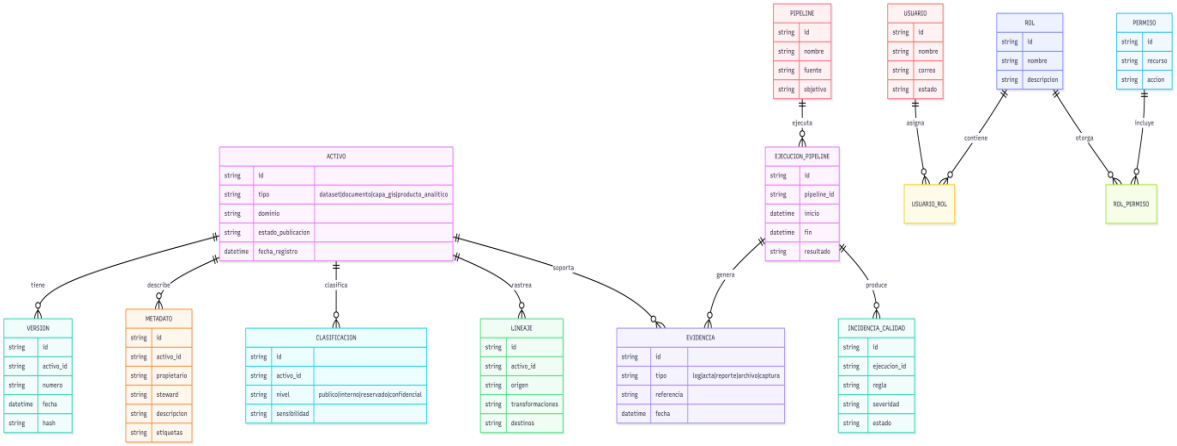
Bitácoras y evidencia:

Cada ejecución registra X-Correlation-ID, timestamps, usuario/rol, resultados de validación, acciones de remediación y evidencia de aceptación en el ambiente correspondiente.



7.3 Repositorio unificado y modelo de datos

El repositorio central soporta almacenamiento para activos estructurados y no estructurados, con políticas de versionamiento, retención y estándares de carga, garantizando integridad, trazabilidad y disponibilidad.



7.4 Catálogo, metadatos, taxonomía y linaje

Se implementa un catálogo de metadatos con requisitos mínimos, taxonomía controlada y linaje extremo a extremo. Se habilitan APIs para consulta de metadatos, taxonomía y linaje y exportaciones compatibles con estándares aplicables.

7.5 Búsqueda e indexación (full-text / vectorial) y NLQ/RAG

Se incorpora un motor de búsqueda unificada con indexación textual y capacidades semánticas para mejorar recuperación de información y soportar consultas en lenguaje natural cuando aplique.

Búsqueda e indexación (full-text / vectorial) y NLQ/RAG – detalle ampliado.

Se implementa un esquema de indexación híbrida para optimizar recuperación de información técnica y documental:

Full-text: indexación de campos y contenido (documentos y metadatos) con facetas por dominio, fecha, geografía, clasificación y estado de publicación.

Vectorial: embeddings sobre contenido autorizado para recuperación semántica, soportando consultas en lenguaje natural.

NLQ/RAG (Retrieval-Augmented Generation) estricto:

El asistente consulta primero el índice/catálogo, recupera evidencia (fragmentos y metadatos) y construye una respuesta citando la fuente del repositorio, evitando alucinaciones mediante umbrales de relevancia, límites de contexto y reglas de autorización por rol.

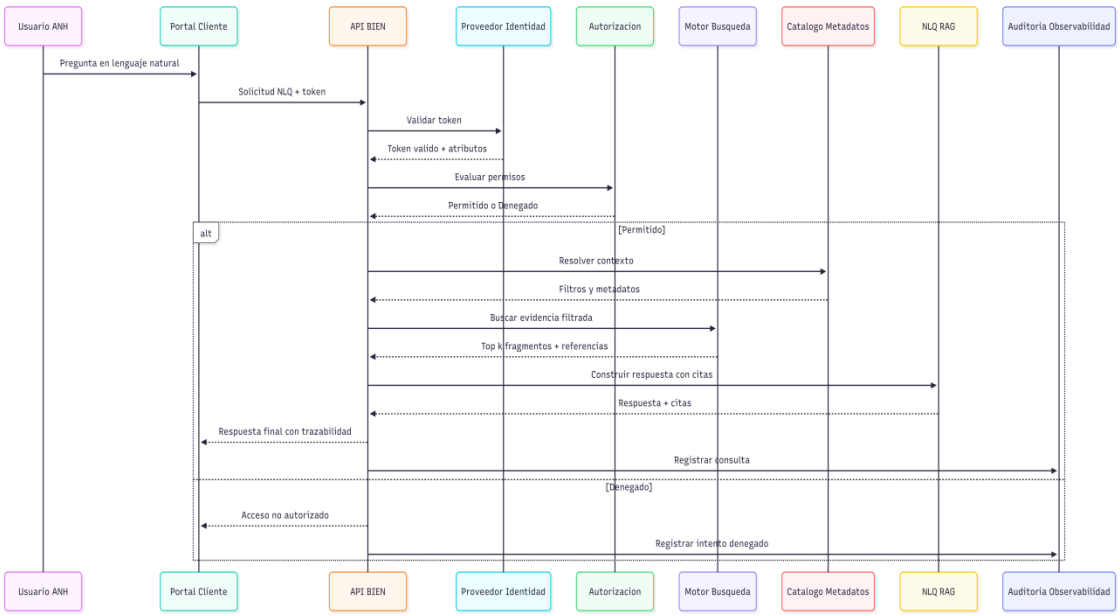
Controles clave:

Autorización antes de recuperar: el motor filtra por permisos (RBAC/ABAC) para impedir exposición por inferencia.

CONFIDENCIAL

Observabilidad: se registran consultas, latencias, top-k, fuentes recuperadas y métricas de calidad de respuesta para mejora continua.

Calidad de búsqueda: se definieron pruebas de precisión/recall por dominios priorizados y validación con usuarios clave.



7.6 GIS: geovisor y servicios geoespaciales

Se integra visualización geográfica mediante geovisor y servicios OGC para capas temáticas, habilitando consulta y consumo por criterios espaciales.

7.7 Calidad, pruebas y desempeño

Se establece trazabilidad REQ-Diseño-Prueba, ejecución de pruebas por ambiente y criterios de aceptación, con evidencias para liberaciones controladas.

8. GESTIÓN DE DATOS Y GOBERNANZA

La gobernanza de datos se implementa mediante políticas, roles y procesos que aseguran consistencia a lo largo del ciclo de vida de la información, incluyendo metadatos mínimos, taxonomía controlada, linaje, retención y auditoría.

8.1 Metadatos mínimos y clasificación

Todo activo publicado debe contar con identificación, propietario, versión, fecha, clasificación y sensibilidad, habilitando control y trazabilidad.

8.2 Linaje y retención

El linaje documenta origen, transformaciones, cargas y consumos. La retención/versionamiento se aplica automáticamente y con auditoría.

9. GESTION DE DATOS Y DECISIONES DE DISEÑO

Modelo de Datos Documental OSDU – SQL Server + MongoDB

Contexto y definiciones

Este proyecto gestiona **información crítica de activos, contratos y documentos**, con necesidades claras y distintas:

- Transacciones confiables
- Integridad referencial
- Auditoría y trazabilidad
- Alto volumen de lectura
- Búsqueda avanzada y semántica

CONFIDENCIAL

- Escalabilidad futura (IA / NLP)

Un solo motor de base de datos **no cubre eficientemente todos estos requerimientos**.

Decisión Clave de Arquitectura

Se optó por una **arquitectura híbrida**, donde:

- SQL Server es el sistema transaccional principal (System of Record)

- MongoDB es la base de datos de consumo (Read / Search Model)

Ambos motores cumplen **roles distintos y complementarios**, no compiten entre sí.

9.1 ¿Por qué SQL Server?

SQL Server fue seleccionado como **fuente principal de datos** por las siguientes razones:

Fortalezas Clave

- Transacciones ACID

- Integridad referencial (PK / FK)

- Consistencia fuerte

- Soporte nativo de CDC

- Capacidades de auditoría y control

Valor para el Proyecto

Garantiza la **calidad y confiabilidad** de la información

CONFIDENCIAL

Centraliza reglas de negocio

Facilita cumplimiento normativo y auditorías

Reduce riesgos operativos

9.2. ¿Por qué MongoDB?

MongoDB fue elegido como **plataforma de consumo** por sus ventajas en escenarios de lectura avanzada:

Fortalezas Clave

- Modelo documental flexible
- Excelente rendimiento en lectura
- Índices avanzados
- Escalabilidad horizontal
- Ideal para búsqueda semántica y NLP

Valor para el Proyecto

- Permite construir documentos enriquecidos por activo
- Reduce complejidad de joins
- Acelera búsquedas y exploración de datos
- Facilita integración con motores de IA

9.3 Sincronización entre Plataformas

La integración se realiza mediante **Change Data Capture (CDC)**:

CONFIDENCIAL

- Cambios capturados desde SQL Server

- Replicación incremental

- Sin impacto en operación productiva

- Consistencia eventual controlada

Esto garantiza:

- Datos actualizados

- Trazabilidad completa

- Posibilidad de reprocesar información

9.4 . Decisiones arquitectónicas

A continuación, se describen las decisiones de diseño adoptadas para la construcción del Modelo de Datos Documental soportado en SQL Server y MongoDB, alineado con los lineamientos de OSDU, DAMA-DMBOK y las necesidades operativas de gestión documental, contratos y activos de información para los dominios de Minas, Pozos y Nuevas Energías.

Estas decisiones buscan garantizar:

- Gobernanza de datos

- Trazabilidad completa

- Escalabilidad

- Auditabilidad

- Separación clara entre metadata estructurada y capacidad de búsqueda

Separación de responsabilidades entre SQL Server y MongoDB

Decisión

CONFIDENCIAL

Se adoptó una arquitectura híbrida donde:

- SQL Server es el sistema de registro (System of Record) de toda la metadata.

- MongoDB actúa como motor indexado de búsqueda, optimizado para consultas rápidas.

Justificación

- OSDU recomienda separar almacenamiento estructurado de mecanismos de indexación.

- DAMA establece que la metadata gobernada debe residir en motores relacionales.

- MongoDB permite búsquedas flexibles (texto, filtros, facetas) sin impactar el OLTP.

Resultado

- SQL Server mantiene integridad referencial, versionamiento y auditoría.

- MongoDB contiene únicamente los campos relevantes para búsqueda.

Contexto del Análisis

Después del análisis de los archivos JSON suministrados (Sondeo de Mercado, CDP, RP, ESET, Personal Mínimo, entre otros) se identificaron patrones consistentes:

Todos los archivos representan documentos asociados a un contrato.

Comparten un conjunto común de atributos técnicos (nombre, tipo, tamaño, checksum, fechas).

Contienen metadata variable según el tipo documental (keywords, theme, spatialCoverage, qualityStatus).

CONFIDENCIAL

Referencian explícitamente la ubicación física del archivo mediante rutas de red (SAN).

Hallazgos relevantes del documento Análisis de Fuentes de Información

El documento Word establece que:

Las fuentes provienen de gestores documentales y sistemas institucionales.

Existe clasificación normativa de documentos por etapas contractuales.

La información debe ser trazable, auditable y reutilizable.

Se requiere interoperabilidad futura con estándares del sector energético.

Decisión de diseño

Decisión – Modelar la carpeta como clasificación documental Las carpetas descritas (1-Sondeo de Mercado, 2-CDP, 3-RP, etc.) no se trataron como simples rutas, sino como clases documentales contractuales.

Esto permite:

Validar obligatoriedad documental.

Homogeneizar clasificación.

Facilitar auditorías.

Uso de ACTIVO_INFORMACION como entidad central Decisión

Se definió ACTIVO_INFORMACION como la entidad núcleo del modelo.

Justificación

- OSDU define el concepto de Data Asset como unidad central.

- Permite desacoplar documentos físicos de su metadata lógica.

- Facilita extensibilidad para diferentes dominios (Minas, Pozos, Energía).

CONFIDENCIAL

Resultado

Todas las entidades clave se relacionan directa o indirectamente con ACTIVO_INFORMACION:

- DOCUMENTO

- METADATA_EXTENDIDA

- POZO / MINA / NUEVA_ENERGIA_ACTIVO

Modelado de dominios: Pozos, Minas y Nuevas Energías

Decisión

Se modelaron entidades específicas:

- POZO

- MINA

- NUEVA_ENERGIA_ACTIVO

Todas relacionadas 1–1 con ACTIVO_INFORMACION.

Justificación

- OSDU separa el tipo de activo del activo genérico.

- Permite agregar atributos especializados sin contaminar el modelo base.

- Facilita evolución futura (ej. sísmica, 2D/3D, encuestas).

CONFIDENCIAL

Resultado

El modelo soporta múltiples dominios manteniendo coherencia estructural.

METADATA_EXTENDIDA como modelo flexible (EAV controlado)

Decisión

Se implementó METADATA_EXTENDIDA bajo un esquema Entidad–Atributo–Valor (EAV) controlado.

Justificación

- Los documentos JSON de entrada contienen metadata variable.

- Evita alterar el modelo físico por cada nuevo atributo.

- OSDU promueve namespaces de metadata extensible.

Resultado

- Se preserva flexibilidad.

- Se mantiene control mediante:

    - namespace

    - atributo

    - valor

-La metadata detallada permanece en SQL Server.

CONFIDENCIAL

Modelo documental contractual

Decisión

Se separaron claramente:

- CONTRATO

- DOCUMENTO

- CLASE_DOCUMENTAL_CONTRATO

Justificación

- La clasificación documental es normativa.

- No debe depender de rutas físicas.

- Facilita búsquedas y cumplimiento legal.

Resultado

- Los documentos se agrupan lógicamente.

- Las carpetas físicas quedan desacopladas del modelo.

Seguridad basada en roles y permisos

Decisión

Se implementó seguridad explícita:

- USUARIO

- ROL

- PERMISO

- ROL_PERMISO

CONFIDENCIAL

- DOCUMENTO_PERMISO

Justificación

- Control de acceso granular.

- Independencia de IAM externo.

- Auditable y extensible.

Resultado

- Permisos por documento.

- Separación de roles y capacidades.

- Soporte a cumplimiento normativo.

Auditoría obligatoria en todas las tablas

Decisión

Todas las tablas incluyen:

- created_at

- created_by

- updated_at

- updated_by

- estado (activo/inactivo)

CONFIDENCIAL

Justificación

- DAMA exige trazabilidad completa.

- No se permiten borrados físicos.

Soporte a auditorías y control interno.

Resultado

- Histórico completo.

- Desactivación lógica.

- Cumplimiento normativo.

No borrado físico (Soft Delete)

Decisión

- No se implementa DELETE físico.

Justificación

- Requisitos legales.

- Trazabilidad.

- OSDU recomienda preservación de activos.

Resultado

- Uso del campo estado.

- Historial intacto.

CONFIDENCIAL

Conclusión

El modelo diseñado:

- Cumple OSDU

- Cumple DAMA-DMBOK

- Es escalable, auditable y gobernable

- Soporta múltiples dominios

- Se integra eficientemente con MongoDB vía CDC

Se anexa Diccionario de Datos

Anexo BIEN_Diccionario_Datos

10. Seguridad, Control y Trazabilidad

La plataforma incorpora controles de acceso por roles y atributos (RBAC/ABAC), gestión de identidades, protección de datos en tránsito y reposo, auditoría de operaciones críticas, observabilidad y prácticas DevSecOps (SAST/DAST/SBOM).

10.1 Auditoría y registros

Se registran eventos de cargas, cambios, accesos, restauraciones e integraciones para soporte de evidencia de cumplimiento.

CONFIDENCIAL

10.2 Observabilidad

Se habilitan logs, métricas y trazas con alertamiento operativo para seguimiento de disponibilidad, desempeño y seguridad.

10.3 DevSecOps

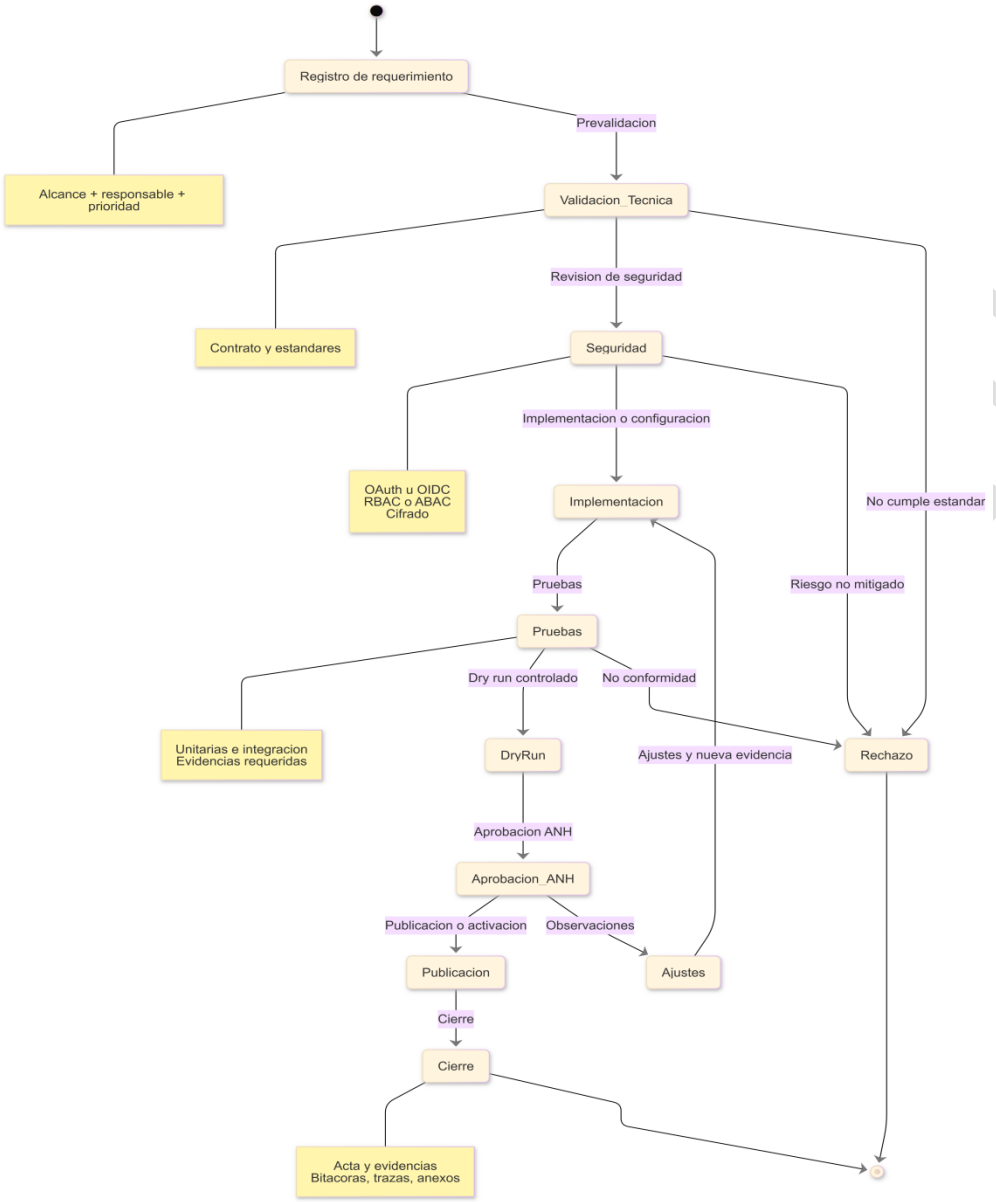
Se integran controles de seguridad en el ciclo de vida: SAST/DAST, SBOM y gestión de secretos, con reportes y remediación documentada.

11. Integraciones e Interoperabilidad

El BIEN expone capacidades de interoperabilidad mediante APIs y servicios OGC. Las integraciones se gobiernan con contratos, versionamiento, autorización y evidencia verificable de cada intercambio.

CONFIDENCIAL

11.1 Workflow de integración y control



11.2 Contratos técnicos

Las integraciones se documentan mediante contratos técnicos (p.ej., REST con idempotencia, timeouts y reintentos) y servicios OGC (WMS/WFS).

11.3 Integración con herramientas especializadas

Se contempla interoperabilidad con herramientas petrotécnicas y componentes GIS/analíticos para intercambio controlado de proyectos, capas y resultados.

11.4 Catálogo de servicios e integración (resumen operativo)

El BIEN expone e integra capacidades mediante servicios versionados, siguiendo contratos técnicos con timeouts, reintentos, idempotencia y trazabilidad (X-Correlation-ID). Las integraciones se habilitan únicamente bajo autorización formal y con evidencia de intercambio.

Servicios internos típicos (agrupados por dominio):

Catálogo/Metadatos: consulta y administración de metadatos, taxonomía, linaje, responsables, estado de publicación.

Búsqueda: consultas unificadas, facetas, guardado de workspaces y exportación controlada.

Ingesta: registro de cargas, validaciones, estado de pipelines y bitácoras.

GIS/OGC: publicación y consumo de capas (WMS/WFS) y consultas espaciales.

Auditoría: consulta de eventos (carga, cambios, accesos, integraciones, restauraciones).

Identidad/Acceso: integración con el proveedor institucional (OIDC/SAML) y gestión de roles/atributos.

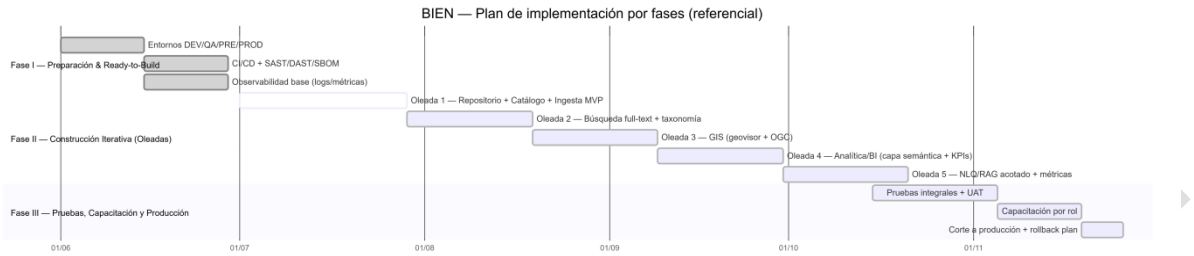
Integraciones externas / mecanismos complementarios:

Además de REST/HTTPS, se soportan mecanismos como SFTP para cargas masivas controladas y webhooks/eventos para notificación de resultados de pipeline y sincronización con sistemas autorizados.

CONFIDENCIAL

12. Estrategia de Implementación y Despliegue

La implementación se ejecuta con enfoque iterativo, ambientes segregados y liberaciones controladas. Se habilita CI/CD, checklist de seguridad y mecanismos de validación Go/No-Go por oleadas.



12.1 Fase I — Preparación & Ready-to-Build

Instalación y validación de entornos DEV/QA/PRE/PROD, habilitación de pipeline, controles SAST/DAST/SBOM y observabilidad base. Construcción de inventario único de fuentes y plantillas de metadatos.

12.2 Fase II — Construcción y Despliegues Iterativos

Construcción modular por oleadas temáticas (datos y catálogo; búsqueda e indexación; GIS; y módulos complementarios), con despliegues en PRE y aprobación Go/No-Go.

12.3 Fase III — Pruebas Integrales, Capacitación y Corte a Producción

Pruebas integrales, capacitación por rol, criterios de aceptación y corte controlado a PROD con mecanismo de rollback y actas de validación.

CONFIDENCIAL

12.4 Gestión por oleadas (MVP e incrementos)

Con base en el plan de implementación, la construcción se organizó por oleadas temáticas, garantizando valor temprano y validación progresiva:

Oleada 1 (MVP): inventario de fuentes, plantillas de metadatos, repositorio inicial, catálogo mínimo y flujo de ingesta controlada.

Oleada 2: búsqueda unificada (full-text) y facetas, consolidación de taxonomía y evidencias de auditoría.

Oleada 3: GIS (geovisor y servicios OGC), publicación de capas y pruebas espaciales.

Oleada 4: analítica/BI (capa semántica, KPIs y tableros oficiales), exportables y control de accesos.

Oleada 5: NLQ/RAG e IA acotada, con controles de autorización, citación y métricas de uso/calidad.

12.5 Controles de liberación (Go/No-Go) y rollback

Cada liberación incluyó checklist de seguridad (SAST/DAST/SBOM), evidencias de pruebas, verificación de observabilidad y un plan de reversión (rollback) con respaldos validados, priorizando continuidad operativa.

13. Consideraciones de Operación y Evolución

La plataforma se diseña para operación continua con monitoreo, gestión de incidentes, mantenimiento y continuidad. Se establecen prácticas de gestión de cambios y versionamiento para evolución segura del sistema.

CONFIDENCIAL

13.1 Escalabilidad y crecimiento

La infraestructura y los componentes se dimensionan para el crecimiento progresivo del volumen de información y demanda, con ajuste de capacidad por métricas.

13.2 Operación temprana

Se contemplan actividades de soporte post-despliegue, verificación de estabilidad, acompañamiento a usuarios clave y seguimiento de adopción.

13.3 Operación, soporte y continuidad

Se definieron procedimientos operativos para: monitoreo (logs/métricas/trazas), gestión de incidentes y problemas, control de cambios, administración de usuarios/roles, y operación de pipelines (reprocesos controlados).

Respaldo, restauración y DRP:

En alineación con RNF-01 (Disponibilidad & DRP) y RT-01 (Infraestructura on-prem), se contemplan respaldos programados, pruebas de restauración, y procedimientos de continuidad que incluyen criterios RPO/RTO acordados, con evidencias de ejecución.

13.4 Evolución y mejoras

El BIEN se concibe como un producto evolutivo: se mantiene un backlog institucional priorizado por valor misional, se versionan contratos de integración, y se gobiernan cambios de taxonomía/metadatos para evitar rupturas en consumo analítico y búsquedas.

CONFIDENCIAL

14. Conclusiones

El desarrollo de la plataforma tecnológica del BIEN materializa un Sistema de Gestión de Datos centralizado, gobernado y seguro, alineado con los pliegos y documentos técnicos. La solución integra ingesta, repositorio, catálogo, búsqueda, GIS e interoperabilidad, con controles de seguridad, trazabilidad y operación para soportar los procesos misionales de la Vicepresidencia Técnica.

Este informe consolida la evidencia técnica del desarrollo e implementación y sirve como insumo para verificación de cumplimiento, operación y evolución futura.

15. Anexos

15.1 Anexo 1 - Infraestructura de datos estructurados no estructurados

15.2 Anexo 2 – Diccionario de datos

15.3 Anexo 3 - Diagrama MER SQLServer

15.4 Anexo 4 - Diagrama MongoDB

16. Glosario

ABAC: Control de acceso basado en atributos (Attribute-Based Access Control). Autoriza según atributos del usuario, recurso y contexto.

API: Interfaz de Programación de Aplicaciones. Conjunto de endpoints para integración y consumo de servicios.

Auditoría: Registro verificable de eventos (cargas, cambios, accesos, integraciones) para evidencia y control.

Backoffice: Interfaz administrativa para gestionar cargas, catálogo, taxonomía, permisos y publicación.

CONFIDENCIAL

BI: Business Intelligence. Capa de analítica y tableros para seguimiento de indicadores.

BIEN: Base Integrada Energética Nacional. Plataforma de gestión, gobierno y consulta de información técnica y documental.

Capa semántica: Modelo de métricas y dimensiones que estandariza definiciones para analítica (KPIs) y reportes.

Catálogo de datos: Inventario gobernado de activos (datasets, documentos, capas) con metadatos, taxonomía y estados de publicación.

CI/CD: Integración y despliegue continuos. Automatización de construcción, pruebas y despliegues por ambiente.

Clasificación de información: Nivel de sensibilidad (público/interno/reservado/confidencial) usado para control de acceso y cumplimiento.

CRS: Sistema de referencia de coordenadas (Coordinate Reference System) para datos geoespaciales.

DAG: Grafo acíclico dirigido. Representa dependencias de tareas en pipelines de datos.

DAST: Dynamic Application Security Testing. Pruebas de seguridad dinámicas sobre aplicaciones en ejecución.

Dataset: Conjunto estructurado de datos (tablas/archivos) con reglas de calidad, metadatos y versiones.

DevSecOps: Integración de seguridad en todo el ciclo de vida de desarrollo y operación.

DRP: Disaster Recovery Plan. Plan de continuidad y recuperación ante desastres.

ELT: Extract–Load–Transform. Patrón donde se carga primero y se transforma posteriormente en el repositorio.

CONFIDENCIAL

ETL: Extract–Transform–Load. Patrón donde se transforma antes de cargar en el repositorio objetivo.

Full-text: Búsqueda textual basada en índices del contenido y metadatos, con ranking y facetas.

GIS: Sistema de Información Geográfica. Componentes para consulta, análisis y visualización espacial.

Hash: Huella digital del contenido usada para validar integridad de archivos/datasets.

Idempotencia: Propiedad de una API donde repetir una operación no genera efectos adicionales no deseados.

Identidad (IdP): Proveedor institucional de autenticación (p.ej., OIDC/SAML) que emite tokens y atributos.

Indexación vectorial: Representación semántica (embeddings) para búsqueda por similitud y recuperación híbrida.

Ingesta: Proceso controlado de incorporación de datos/documentos a la plataforma, con validaciones y evidencia.

Linaje: Trazabilidad de origen, transformaciones y consumos de un activo de información.

Metadatos: Datos que describen otros datos: propietario, dominio, descripción, temporalidad, geografía, etc.

MVP: Producto mínimo viable. Primer incremento funcional que entrega valor temprano con alcance acotado.

NLQ: Natural Language Query. Consulta en lenguaje natural sobre datos y documentos.

OGC: Open Geospatial Consortium. Estándares para servicios geoespaciales.

OKD: Distribución comunitaria de Kubernetes compatible con OpenShift (según contexto de despliegue).

CONFIDENCIAL

Pipeline: Flujo automatizado de procesamiento (carga, validación, transformación, publicación, indexación).

RBAC: Control de acceso basado en roles (Role-Based Access Control).

RAG: Retrieval-Augmented Generation. Respuesta generada apoyada en recuperación de evidencia con citación.

RPO/RTO: Objetivos de recuperación: RPO (pérdida de datos tolerable) y RTO (tiempo máximo de recuperación).

SBOM: Software Bill of Materials. Inventario de componentes de software para gestión de riesgos.

SIEM: Security Information and Event Management. Correlación y monitoreo de eventos de seguridad.

Taxonomía: Clasificación controlada de dominios/temas para organizar activos y habilitar búsqueda y gobierno.

WFS/WMS: Servicios OGC para consulta/visualización de capas geoespaciales (Features/Map Service).

X-Correlation-ID: Identificador único para rastrear una transacción extremo a extremo en logs y auditoría.

ABAC: Control de acceso basado en atributos (Attribute-Based Access Control). Autoriza según atributos del usuario, recurso y contexto.

API: Interfaz de Programación de Aplicaciones. Conjunto de endpoints para integración y consumo de servicios.

Auditoría: Registro verificable de eventos (cargas, cambios, accesos, integraciones) para evidencia y control.

Backoffice: Interfaz administrativa para gestionar cargas, catálogo, taxonomía, permisos y publicación.

BI: Business Intelligence. Capa de analítica y tableros para seguimiento de indicadores.

CONFIDENCIAL

BIEN: Base Integrada Energética Nacional. Plataforma de gestión, gobierno y consulta de información técnica y documental.

Capa semántica: Modelo de métricas y dimensiones que estandariza definiciones para analítica (KPIs) y reportes.

Catálogo de datos: Inventario gobernado de activos (datasets, documentos, capas) con metadatos, taxonomía y estados de publicación.

CI/CD: Integración y despliegue continuos. Automatización de construcción, pruebas y despliegues por ambiente.

Clasificación de información: Nivel de sensibilidad (público/interno/reservado/confidencial) usado para control de acceso y cumplimiento.

CRS: Sistema de referencia de coordenadas (Coordinate Reference System) para datos geoespaciales.

DAG: Grafo acíclico dirigido. Representa dependencias de tareas en pipelines de datos.

DAST: Dynamic Application Security Testing. Pruebas de seguridad dinámicas sobre aplicaciones en ejecución.

Dataset: Conjunto estructurado de datos (tablas/archivos) con reglas de calidad, metadatos y versiones.

DevSecOps: Integración de seguridad en todo el ciclo de vida de desarrollo y operación.

DRP: Disaster Recovery Plan. Plan de continuidad y recuperación ante desastres.

ELT: Extract–Load–Transform. Patrón donde se carga primero y se transforma posteriormente en el repositorio.

ETL: Extract–Transform–Load. Patrón donde se transforma antes de cargar en el repositorio objetivo.

CONFIDENCIAL

Full-text: Búsqueda textual basada en índices del contenido y metadatos, con ranking y facetas.

GIS: Sistema de Información Geográfica. Componentes para consulta, análisis y visualización espacial.

Hash: Huella digital del contenido usada para validar integridad de archivos/datasets.

Idempotencia: Propiedad de una API donde repetir una operación no genera efectos adicionales no deseados.

Identidad (IdP): Proveedor institucional de autenticación (p.ej., OIDC/SAML) que emite tokens y atributos.

Indexación vectorial: Representación semántica (embeddings) para búsqueda por similitud y recuperación híbrida.

Ingesta: Proceso controlado de incorporación de datos/documentos a la plataforma, con validaciones y evidencia.

Linaje: Trazabilidad de origen, transformaciones y consumos de un activo de información.

Metadatos: Datos que describen otros datos: propietario, dominio, descripción, temporalidad, geografía, etc.

MVP: Producto mínimo viable. Primer incremento funcional que entrega valor temprano con alcance acotado.

NLQ: Natural Language Query. Consulta en lenguaje natural sobre datos y documentos.

OGC: Open Geospatial Consortium. Estándares para servicios geoespaciales.

OKD: Distribución comunitaria de Kubernetes compatible con OpenShift (según contexto de despliegue).

Pipeline: Flujo automatizado de procesamiento (carga, validación, transformación, publicación, indexación).

CONFIDENCIAL

RBAC: Control de acceso basado en roles (Role-Based Access Control).

RAG: Retrieval-Augmented Generation. Respuesta generada apoyada en recuperación de evidencia con citación.

RPO/RTO: Objetivos de recuperación: RPO (pérdida de datos tolerable) y RTO (tiempo máximo de recuperación).

SBOM: Software Bill of Materials. Inventario de componentes de software para gestión de riesgos.

SIEM: Security Information and Event Management. Correlación y monitoreo de eventos de seguridad.

Taxonomía: Clasificación controlada de dominios/temas para organizar activos y habilitar búsqueda y gobierno.

WFS/WMS: Servicios OGC para consulta/visualización de capas geoespaciales (Features/Map Service).

X-Correlation-ID: Identificador único para rastrear una transacción extremo a extremo en logs y auditoría.

CONFIDENCIAL