SECCION SITIOS

Esta sección permite gestionar los sitios web para realizar scraping, activación o desactivación y definir las frecuencias

- SITIOS

SITIOS

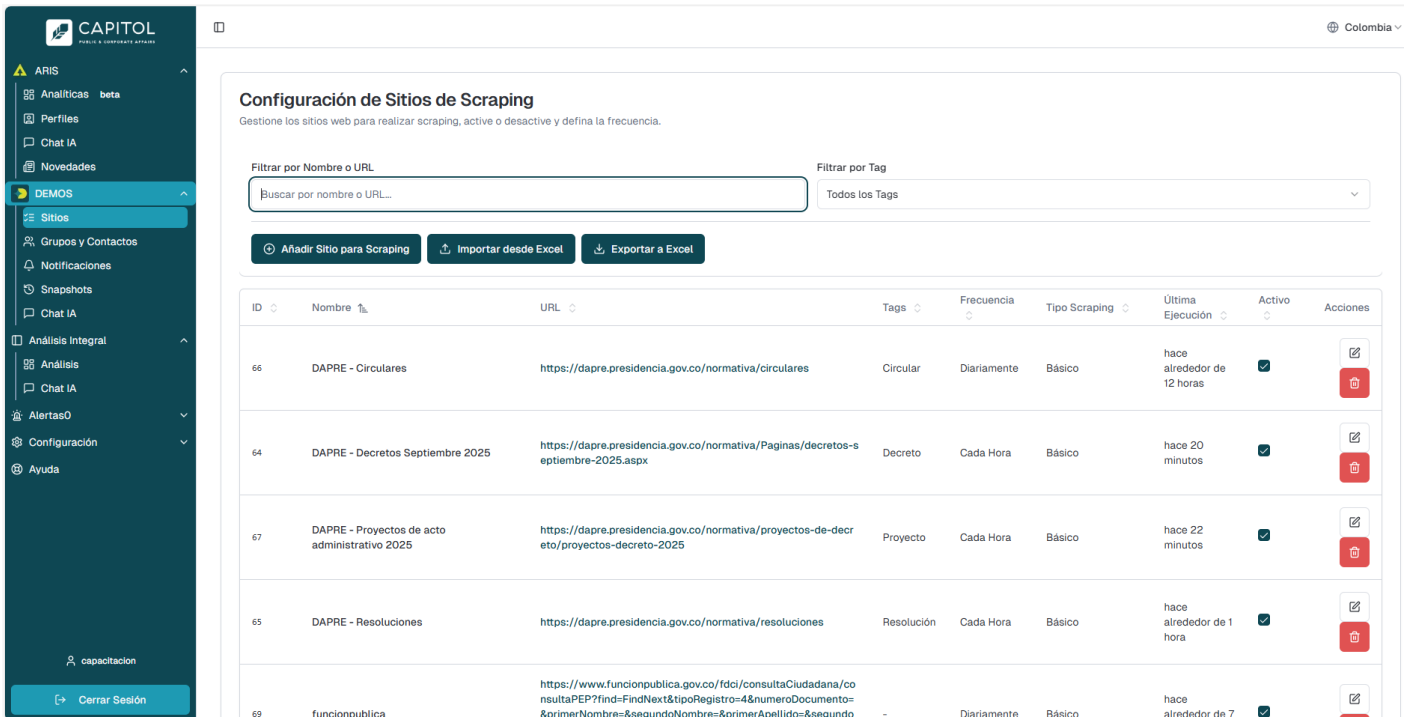
Sección Sitios

Esta sección permite gestionar los sitios web para realizar scraping, activación o desactivación y definir las frecuencias.

- Configurar los sitios web que deseas monitorear. Define la URL, la frecuencia de revisión y el tipo de monitoreo (básico, PDF, etc.)..
- Exportar e importar en excel los datos de scraping.
- Editar y/o eliminar la configuración de los sitios.

Podemos ubicar la información en una tabla la cual está conformado por las siguientes variables: ID, Nombre, Redes sociales, Caracterización, Notificar, Ultima sincronización, Acciones.

- **ID**: Identificador generado de manera automática para los perfiles creados.
- **Nombre:** Corresponde al nombre que se le asigna al sitio.
- **Url:** URL del sitio.
- **Tags:** Tags o clasificación que se asigna al sitio, importante separar por comas.
- **Frecuencia:** Frecuencia en la que se realizara el scraping.
- **Tipo Scraping:** Selección del tipo de scraping; básico, PDF scanning, personalizado (Endpoint Rest) personalizado (Script playwright).
- **Ultima ejecución:** Fecha de la ultima ejecución del scraping.
- **Activo:** Check de marcación de estado.
- **Acciones:** Acción de editar y /o eliminar.

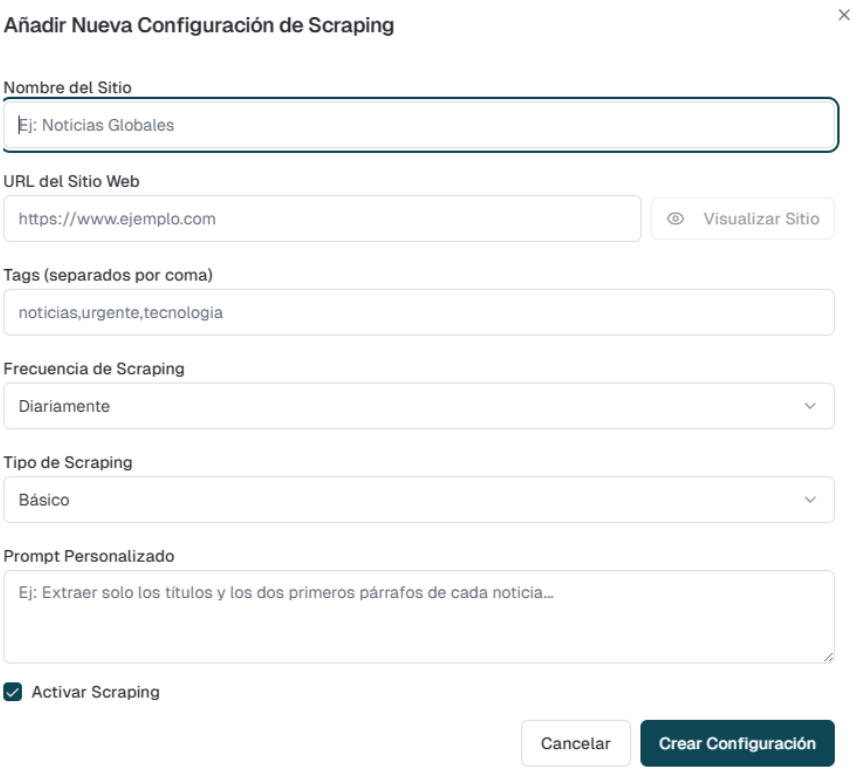


Añadir sitio para scraping

Esta funcionalidad permite añadir nuevos sitios para realizar el scraping. Los campos a diligenciar son:

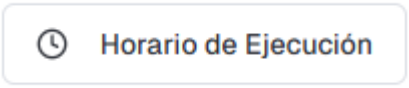
- **Nombre del sitio:** Indicar el nombre asignar al sitio, nombre descriptivo para identificarlo facilmente: ("ministerio de salud").
- **URL del sitio Web:** Indicar la URL del sitio a monitorear.
- **Tags:** Indicar el o los tags asignar al sitio, se debe tener en cuenta que deben separare por coma, palabra clave para categorizar el sitio: ("gobierno, salud).
- **Frecuencia de scraping:** Frecuencia en que se realizara el scraping:
    - Diariamente
    - Semanalmente
    - Mensualmente
    - Cada hora
    - Cada 30 minutos
    - Cada 15 minutos
    - Cada 10 minutos
    - Cada 5 minuto

- **Tipo de scraping:** Indicar el tipo de scraping (método extracción de datos):
    - **Básico:** Monitoreo estándar del contenido HTML. Ideal para sitios simples
    - **PDF scanning:** Descarga y analiza el texto de los documentos PDF en la página. Puede activar la opción de OCR para extraer texto de imágenes dentro de los PDFs.
    - **PDF Crawling:** Similar a PDF Scanning, pero además recorre los enlaces internos del sitio para encontrar más PDFs en páginas secundarias.
    - **Personalizado (Endpoint Rest):** Para sitios complejos que requieren un servicio externo (API) para obtener los datos. Debe proporcionar la URL del endpoint.
    - **Personalizado (Script playwright):** Escriba un script personalizado en Playwright/JavaScript para navegar e interactuar con sitios dinámicos y extraer la información precisa. Puede probar su script directamente desde el formulario
    - **Prompt Personalizado:** (Opcional) Proporcione instrucciones específicas para la IA sobre cómo debe analizar el contenido extraído (ej: "Extraer solo los títulos y los dos primeros párrafos de cada noticia").
    - **Activar Scraping:** Marque esta casilla para que el monitoreo de este sitio esté activo.



Pantalla de añadir nueva configuración de scraping.

- **Horario de Ejecución:** Configure un horario global (días y rango de horas) en el que se ejecutarán todos los monitoreos



Importar desde Excel

Esta funcionalidad permite importar archivos en formato Excel de manera masiva.

Exportar desde Excel

Esta funcionalidad permite exportar y/o descargar archivos en formato Excel.