Implementación de Anubis WAF en el SID

Implementación de Anubis WAF en el SID: Protegiendo la Infraestructura de la UNCuyo

Artículo para el blog del SID — sid.uncu.edu.ar

Introducción

En el **Sistema Integrado de Gestión de Bibliotecas (SID)** de la Universidad Nacional de Cuyo, la infraestructura digital es un activo crítico. Miles de investigadores, estudiantes y docentes acceden diariamente a recursos académicos a través de más de 20 portales web hospedados bajo los dominios `bdigital.uncu.edu.ar`, `revistas.uncu.edu.ar`, `sid.uncu.edu.ar`, entre otros.

Esta exposición constante atrae no solo usuarios legítimos, sino también **bots automatizados**, raspadores de contenido y herramientas de inteligencia artificial que, sin control, pueden degradar el servicio, saturar los servidores y acceder a datos de manera no autorizada.

Frente a este escenario, el SID implementó **Anubis WAF** (Web Application Firewall) como capa de protección activa. Este artículo documenta la primera prueba piloto, sus resultados y la importancia de esta decisión para la comunidad universitaria.

¿Qué es Anubis WAF?

Anubis es un firewall de aplicaciones web de código abierto diseñado para analizar tráfico HTTP/HTTPS en tiempo real, identificar patrones de comportamiento automatizado y aplicar reglas de bloqueo o desafíos (challenges) sin interrumpir el acceso de usuarios legítimos.

A diferencia de los firewalls tradicionales que trabajan en capa de red, Anubis opera en capa de aplicación, lo que le permite inspeccionar:

- **Cabeceras User-Agent** — Identificación del cliente que realiza la petición.
- **Patrones de navegación** — Frecuencia, orden y tipo de rutas accedidas.
- **Comportamiento de bots** — Scrapers, crawlers no autorizados, herramientas de raspado de contenido, prompts de IA.
- **Reputación de IP** — Listas de IPs conocidas por actividad maliciosa.

¿Por qué es necesario proteger la infraestructura?

1. Protección del Hardware y los Recursos de Red

Cada petición HTTP consume recursos del servidor: CPU, memoria, conexiones de red y ancho de banda. Un bot mal diseñado o excesivamente agresivo puede generar miles de solicitudes por minuto, saturando los servidores y afectando la experiencia de los usuarios legítimos.

En el SID, donde el tráfico se distribuye entre más de 20 instancias y portales, un evento de este tipo podría dejar fuera de servicio a toda la red de bibliotecas universitarias.

2. Protección de Datos y Contenido Académico

El contenido almacenado en el SID — tesis, revistas científicas, documentos históricos, bases de datos especializadas — es propiedad intelectual generada por la comunidad universitaria. Los raspadores automatizados pueden copiar este contenido masivamente para alimentar modelos de lenguaje o bases de datos comerciales, sin autorización ni atribución.

Proteger estos datos no es solo una cuestión técnica: es una responsabilidad institucional.

3. Continuidad del Servicio

Un sistema de salud digital como el SID debe estar disponible cuando los investigadores lo necesitan. El bloqueo proactivo de bots asegura que los servidores mantengan capacidad para atender a usuarios reales, especialmente en períodos de alta demanda (inicio de cuatrimestre, época de exámenes, entregas de trabajos finales).

4. Cumplimiento y Gobernanza de Datos

La UNCuyo, como institución pública, tiene la obligación de garantizar la disponibilidad, integridad y confidencialidad de sus datos digitales. Implementar un WAF como Anubis forma parte de una estrategia de gobernanza de datos responsable y alineada con las mejores prácticas de ciberseguridad.

Resultados de la Primera Prueba Piloto

La primera implementación de Anubis en el SID se realizó sobre el archivo de logs del **2 de junio de 2026**, con un volumen total de **1.34 GB** de eventos registrados.

Estadísticas Generales

Métrica	Valor
Total de bloqueos (DENY)	492,425
Total de desafíos (CHALLENGE/CAPTCHA)	53,287
Motivo principal de bloqueo	bot/block-ai-scrapers (100%)

Acciones Aplicadas

Acción	Descripción	Cantidad
DENY	Bloqueo directo de la petición	492,425
CHALLENGE	Verificación CAPTCHA antes de permitir acceso	1,698,332

Principales Rastreadores Bloqueados

User Agent	Porcentaje del Total Bloqueado
SemrushBot	44.3%
meta-externalagent (Facebook/Meta AI)	33.7%
generic-browser	8.1%
Baiduspider	4.7%
YandexBot	1.6%
AhrefsBot	0.6%
OJS-Research-Scraper	0.6%
Sogou web spider	0.9%

> **Dato notable:** Más del 78% del tráfico bloqueado corresponde a bots de empresas de SEO, raspadores de IA y motores de búsqueda que no respetan los protocolos de exclusión (`robots.txt`) o que acceden con patrones de comportamiento automatizado.

Rutas Más Afectadas

Las rutas que recibieron mayor cantidad de intentos bloqueados fueron (pertenecen a distintos dominios y aplicaciones):

Ruta	Peticiones Bloqueadas
`/explorador3/MyResearch/SaveSearch`	151,127
`/buscador.php`	134,434
`/form_contacto/`	30,723
`/app/navegador/`	25,340
`/fichas.php`	22,466

Estas rutas, que exponen funcionalidades de búsqueda y consulta de catálogos, son objetivos frecuentes de herramientas de raspado masivo que buscan indexar o extraer el contenido del repositorio.

Distribución por Host

El piloto cubrió múltiples portales del SID. A continuación, los más relevantes:

Host	DENY	CHALLENGE
bibliotecas.uncuyo.edu.ar	166,309	653,090
bdigital.uncuyo.edu.ar	45,886	189,574
bdigital.uncu.edu.ar	24,158	36,424
revistas.uncu.edu.ar	26,855	187,014
economicas.bdigital.uncu.edu.ar	15,915	27,598

entre otros

Casos de Estudio: ¿Qué fue bloqueado?

1. SemrushBot (44.3% del tráfico bloqueado)

Semrush es una plataforma de análisis SEO utilizada por profesionales de marketing digital. Su bot rastrea sitios web para construir bases de datos de backlinks y posicionamiento. En el contexto del SID, este bot accede masivamente a las rutas del catálogo sin aportar valor a la comunidad universitaria, consumiendo recursos que deberían estar disponibles para los usuarios reales.

2. meta-externalagent (33.7%)

Este agente corresponde a herramientas de inteligencia artificial de Meta (Facebook) que entrenan modelos de lenguaje con contenido de la web. Acceden a los portales del SID para extraer datos textuales con fines comerciales, sin autorización institucional.

3. Otros bots de IA y raspadores

Bots de Baidu, Yandex, Ahrefs, Sogou y scrapers genéricos completan el panorama. Si bien algunos de estos motores de búsqueda pueden tener valor para la visibilidad del SID, su comportamiento automatizado sin control puede generar carga excesiva en los servidores.

Conclusiones y Próximos Pasos

La primera implementación de Anubis WAF en el SID demuestra que **la proporción de tráfico automatizado no autorizado es significativamente mayor al tráfico legítimo** en los portales de la universidad. Bloquear este tráfico protege los recursos de infraestructura y garantiza que los servidores respondan eficientemente a quienes realmente necesitan acceder a la información académica.

Beneficios observados:

- □ Reducción directa del tráfico no autorizado hacia los servidores del SID.
- □ Preservación de ancho de banda y recursos de cómputo para usuarios legítimos.
- □ Protección del contenido académico contra raspado masivo.
- □ Registro detallado para auditoría y análisis de patrones de acceso.

Líneas de trabajo futuras:

- □□ **Refinamiento de reglas:** Ajustar los thresholds para minimizar falsos positivos sin comprometer la seguridad.
- □□ **White-listing inteligente:** Permitir acceso sin desafíos a bots de búsqueda legítimos que respeten `robots.txt`.
- □□ **Monitoreo en tiempo real:** Integrar los logs de Anubis con los dashboards existentes del SID.
- □□ **Comunicación a la comunidad:** Informar a investigadores e instituciones aliadas sobre el acceso legítimo automatizado y cómo solicitar excepciones si es necesario.

Acceso al Dashboard

El dashboard público de análisis de logs de Anubis está disponible en:

https://bibliotecas.uncuyo.edu.ar/tools/anubis/

Este panel se actualiza con cada nuevo análisis de logs y permite visualizar en tiempo real la actividad bloqueada, los patrones de tráfico y la distribución por host.

Implementado por el equipo de infraestructura del SID — UNCuyo. Para consultas o solicitudes de whitelisting, contactar a: horacio.degiorgi@uncuyo.edu.ar

---