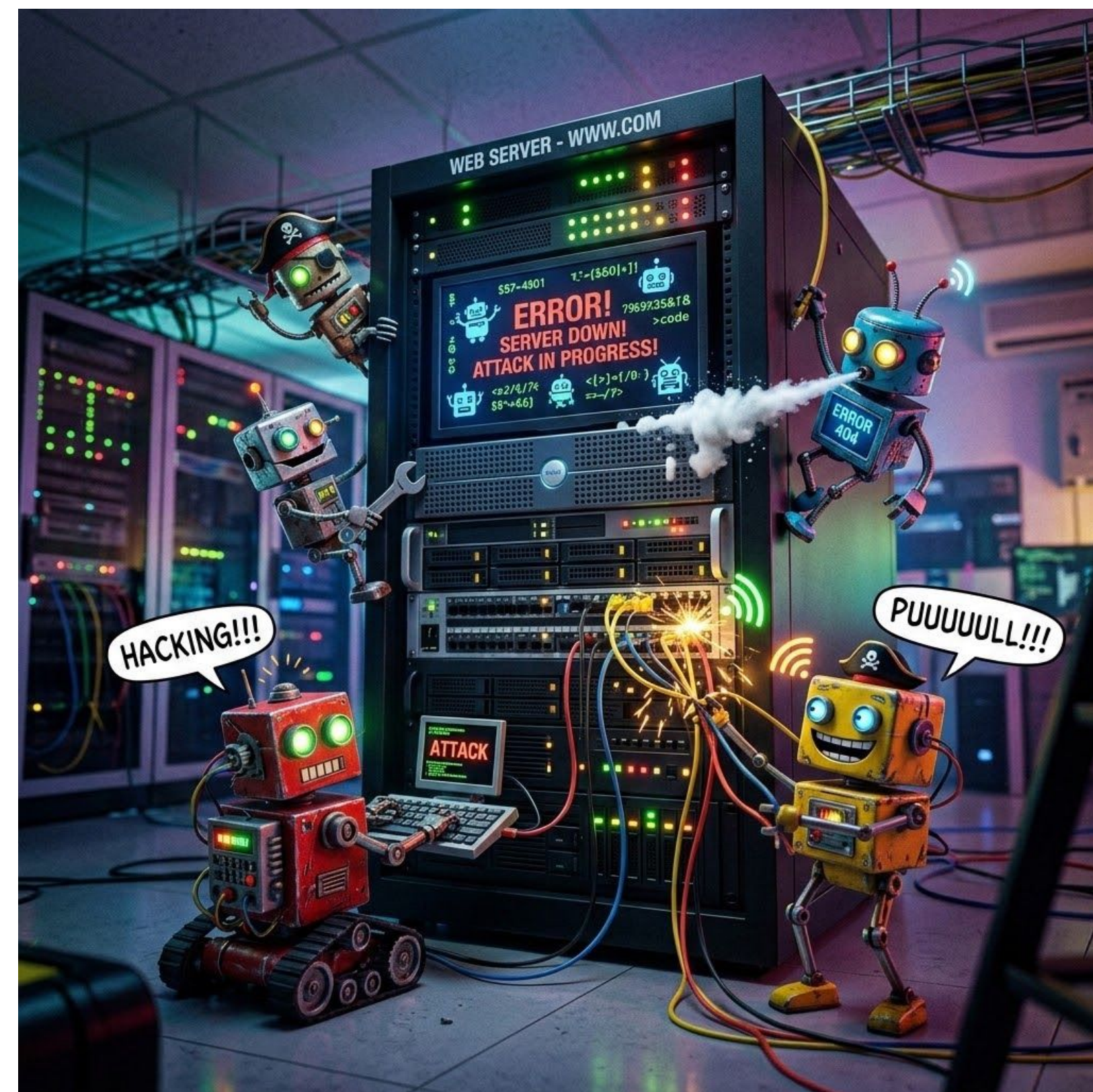


Mitigación de Bots, nuevos desafíos en la ciencia abierta y las bibliotecas.

Horacio Degiorgi

(UNCuyo - LaReferencia)



Agenda de Temas

01

Activos a proteger

Cuáles son nuestros activos a proteger. Identificación de los recursos, datos e información crítica que debemos resguardar.

02

El enemigo

Identificando al enemigo. Análisis de amenazas, automatizaciones maliciosas y patrones de comportamiento de bots.

03

Estrategia

Cómo tenemos que proteger. Definición de protocolos de seguridad, mitigación y mejores prácticas.

04

Impacto

Cómo cambia esto nuestras aplicaciones. Transformación arquitectónica y adaptación de los sistemas existentes.

Activos a proteger



Componentes y Protección

Las aplicaciones tienen distintos componentes que deben ser protegidos y a la vez expuestos.

- La infraestructura
- Los datos
- La red
- Los usuarios
- Los recursos de información



Usabilidad y Rendimiento

Durante los últimos años nos hemos preocupado de que las aplicaciones tengan una gran usabilidad, muchos puntos de acceso, una apertura completa, que respondan rápido y otras funcionalidades.

- Más links
- Mejores buscadores
- Más velocidad
- Clientes humanos



Objetivo Principal

Dar visibilidad a nuestros recursos de información: tanto los generados por nosotros como los ofrecidos a nuestra audiencia.

- Acceso sin restricciones
- Modelo de acceso abierto
- Webs preparadas para alta demanda "humana"
- Apertura especial para los crawlers buenos.

Visitantes deseados e intrusos



Visita Web Normal

Navegación estándar realizada por un usuario humano a través de un navegador web interactuando con la interfaz.

Características

- Interacción humana manual
- Ejecución completa de JS/CSS
- Patrón de tráfico **orgánico**



Crawling

Proceso automatizado de rastreo e indexación de enlaces web para descubrir y mapear la estructura del sitio.

Características

- Rastreo masivo de enlaces (URLs)
- Usado por buscadores (Googlebot)
- Sigue estructura hipertextual

Respetan reglas antiguas :
Robots.txt



Scraping

Extracción automatizada y localizada de datos específicos desde páginas web para su almacenamiento o análisis.

Características

- Extracción dirigida de contenidos
- Procesamiento de HTML/APIs
- Genera alta carga en servidor

Generalmente no respetan reglas y cometen **abusos** en base a pruebas de fuerza bruta.

Componentes de una Página Web | Ejemplo: Catálogo de bibliotecas



Consultas Directas

Elementos interactivos que generan peticiones en tiempo real a la base de datos:

- Cuadros de búsqueda con autocompletado
- Personalizaciones por usuario



Filtros y Navegación

Herramientas de refinamiento para exploración del catálogo:

- Las facetas
- Los filtros
- Los ordenamientos



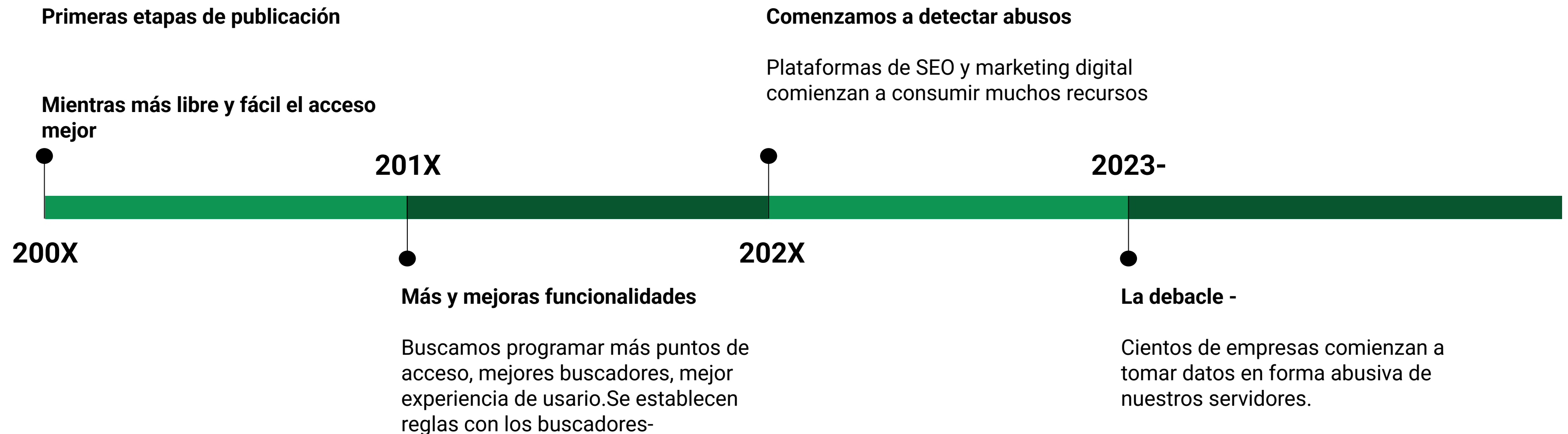
Consumo de Recursos

Impacto técnico generado al interactuar con el sistema:

- **Todas producen consumo de recursos al ser navegadas**
- Procesamiento continuo en servidor y base de datos

Las webs las construimos para seres humanos.
Dejamos puertas especiales para los crawlers y spiders.
Actualmente las reglas y acuerdos son vulnerados por empresas y por particulares

Ciclo de vida de las webs | Bibliotecas, revistas en línea y repositorios institucionales



¿Qué proteger?



Repositorios Institucionales

Documentación académica, tesis, producción científica y memoria digital de la institución.



Catálogos de Biblioteca

Sistemas OPAC, registros bibliográficos, bases de datos de inventario y consulta pública.



Artículos de Revistas

Publicaciones seriadas, investigaciones arbitradas e índices de acceso abierto o restringido.



Archivos de Imágenes

Acervos fotográficos, colecciones históricas digitalizadas, mapas y recursos multimedia.



Documentos Preprint

Borradores científicos preliminares, informes técnicos y avances de investigación no editados.

Formas de protección | Acceso Abierto / Libre (Sin Mitigación Activa)



Enfoque y Funcionamiento

Enrutamiento directo al backend sin barreras de filtrado de bots.

El tráfico llega sin inspección previa ni desafíos PoW, dependiendo solo de configuraciones estándar del servidor web (timeouts básicos).



Ventajas

- Cero fricción para cualquier tipo de integrador externo, agregadores de datos o herramientas legítimas.
- Facilidad absoluta de depuración y desarrollo.



Desventajas

- Exposición completa a scraping agresivo y abuso de recursos.
- Vulnerabilidad alta a ataques DDoS en capa de aplicación (L7).



Casos de Uso Ideales (2026-)

- Entornos locales y staging.
- Feeds RSS públicos.
- Endpoints de Webhooks.
- Servicios internos detrás de VPN.

Formas de protección | Protección Reputacional y de Comportamiento (Solo CrowdSec)



Enfoque y Funcionamiento

Detección: Logs, heurística de red e inteligencia colectiva (*Threat Intelligence*).

- Monitoriza logs web (Nginx, Traefik, Apache) y red.
- Baneo automático vía Firewall/Bouncer ante excesos de umbral.
- Alimentado por lista comunitaria global.



Ventajas

- Totalmente transparente para usuarios legítimos (cero latencia de desafíos).
- Compatible de forma nativa con APIs y aplicaciones móviles.
- Muy bajo consumo de recursos en el servidor.



Desventajas

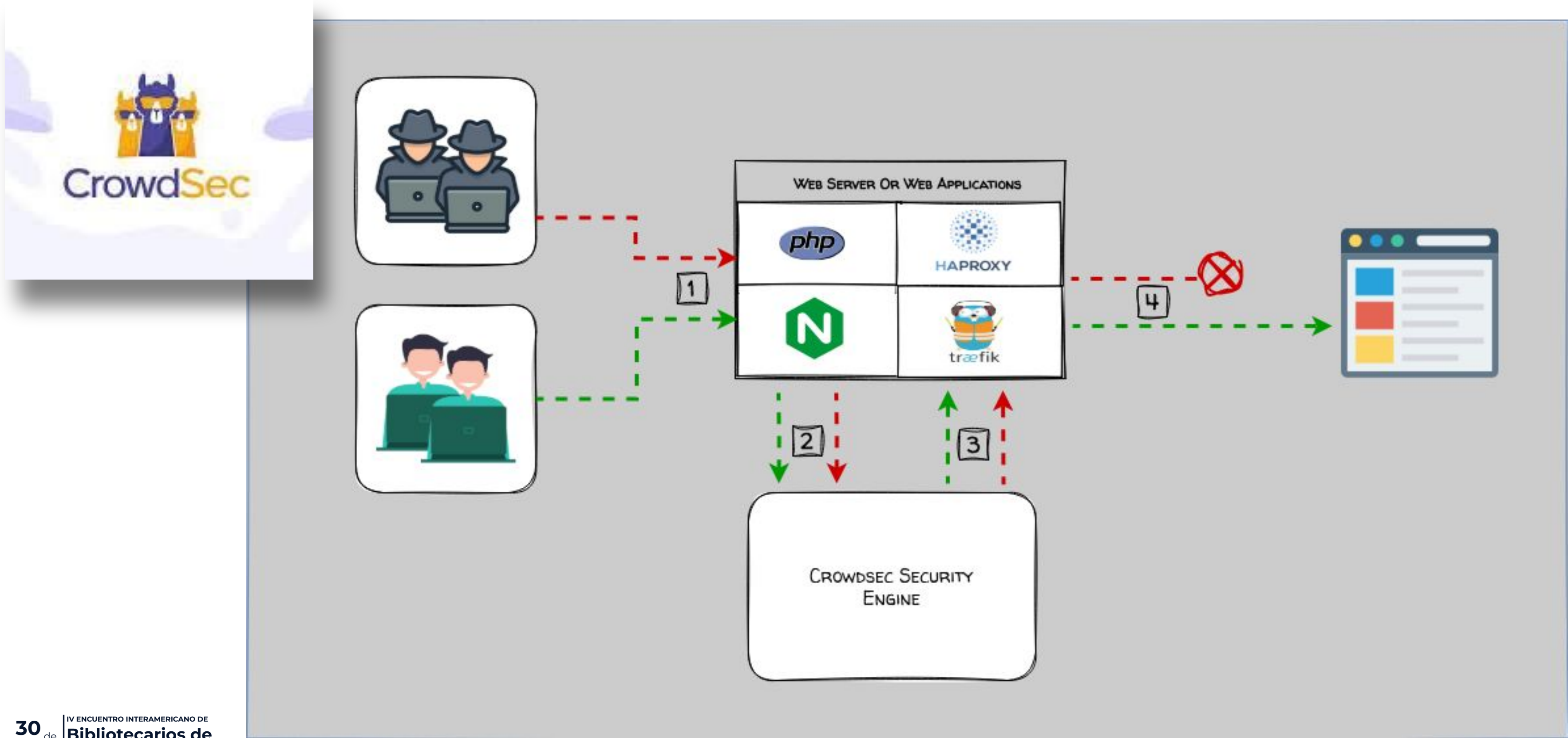
- Vulnerable a bots "lentos y distribuidos" (*low and slow*).
- Ataques desde miles de IPs residenciales que evitan activar umbrales de frecuencia.

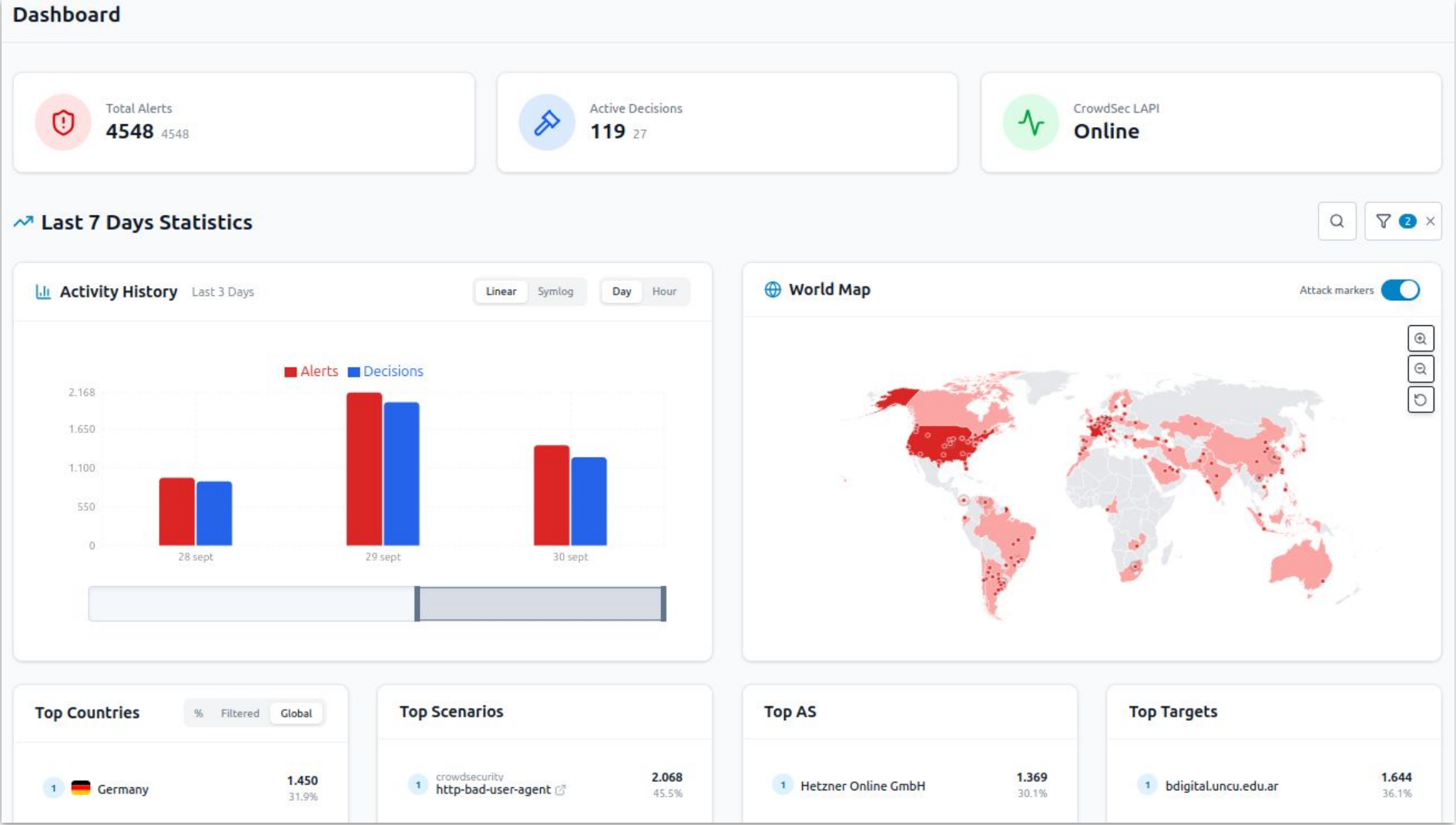


Casos de Uso Ideales

- Sitios institucionales.
- APIs públicas con autenticación previa.
- Entornos donde la experiencia de usuario y velocidad de carga son críticas.







Alert Details #55885

Captured at 30/9/2026, 07:52:19

SCENARIO crowdsecurity http-crawl-non_statics	LOCATION Singapore Lat: 1.314, Long: 103.844	IP / RANGE 146.190.107.21 Range: 146.190.0.0/17 DIGITALOCEAN-ASN (AS14061)	TARGET docs2.a
---	--	---	-------------------

Ip 146.190.107.21 performed 'crowdsecurity/http-crawl-non_statics' (140 events over 51.389382585s) at 2026-09-30 22:52:19.315215932

Decisions Taken

ID	TYPE	VALUE	EXPIRATION	ORIGIN	VIEW
#4493429	ban	146.190.107.21	1h40m43s	crowdsec	

Context

status	301
user_agent	Mozilla/5.0 (Linux; Android 10; K) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/114.0.0.0 Mobile
method	GET
target_uri	/lx.php /mad.php /mail.php /marijuana.php /mar.php Show all 6 values (1 more)

Nivel 1 **Máxima Protección (CrowdSec + Anubis)**

Enfoque: Defensa en profundidad (Reputación IP + Desafío criptográfico/PoW en capa de aplicación)

Filtro Perimetral (CrowdSec)

Bloquea de inmediato direcciones IP reportadas globalmente por actividad maliciosa o patrones abusivos locales (fuerza bruta, escaneo de puertos, spam).

Filtro de Comportamiento y Cliente (Anubis)

Para el tráfico que supera el primer filtro, Anubis actúa como proxy/sidecar evaluando firmas y aplicando desafíos de prueba de trabajo (*Proof-of-Work* / JavaScript silencioso). Si el cliente no resuelve el desafío, la petición nunca llega a la aplicación.

✓ **Ventajas**

- Mitiga scrapers avanzados, bots headless y crawlers de IA que rotan IPs residenciales.
- Reduce a cero la carga de bots pesados en la base de datos y backend.
- Permite separar usuarios “reales” de otros nocivos.

⚠ **Desventajas**

- Mayor consumo de CPU del cliente (por PoW) y posible fricción en clientes sin JavaScript o APIs no exentas.

🎯 **Casos de Uso Ideales**

- Formularios sensibles (login, registro).
- Plataformas con contenido valioso susceptible de scraping masivo.
- Servicios bajo ataque recurrente.

<https://bibliotecas.uncuyo.edu.ar/cdhexplorador/>



¿Dónde los estamos usando?



Biblioteca Digital Repositorio Institucional de la UNCUYO. Encontrá tesis de grado y posgrado, revistas científicas, informes de investigación, eventos, congresos, audios y videos. acceder	Sistema Integral de Gestión Aplicación de Gestión de los recursos de Bibliotecas UNCUYO acceder	Explorador de recursos Explorador de recursos de información disponibles para toda la comunidad universitaria. acceder	Revistas UNCUYO Portal que reúne revistas digitales en Open Journal System (OJS) publicadas por la Universidad Nacional de Cuyo y gestionadas por el Sistema Integrado de Documentación. acceder
Explorador CDH Explorador de registros de Documentos históricos de la UNCUYO o de Instituciones, Personas o Familias relacionados con la misma. acceder	Digesto Administrativo Compendio ordenado y sistematizado de todas las normas emanadas de el/la Rector/a, el Consejo Superior y la Asamblea Universitaria de la Universidad Nacional de Cuyo. acceder	Listado de DOIs Identificadores de objetos digitales DOI, asignados por la UNCUYO desde el 2021, como parte de los procesos que se desarrollan en la Biblioteca Digital. acceder	Micrositios Micrositios de la Biblioteca Digital creados para agrupar objetos digitales de distintas temáticas, áreas o dependencias de la UNCUYO. acceder
Portal REA Objetos de aprendizaje que han resultados de trabajos de investigación o de actividades académicas de los distintos niveles de la universidad. acceder	Colección Digital Lagiglia Fondo documental de Dr. Humberto Lagiglia. El registro se compone, entre otros documentos, de investigaciones en curso y desarrolladas, estudios, artículos, material fotográfico, libretas de campo, entre otros. acceder		

¿Cómo seguir?

¿ Protegemos o compramos más hardware ?

Iniciativas Mundiales para re-establecer reglas de uso y no abuso

Ofrecer alternativas de contenido



The /llms.txt file, v2

A proposal to standardise on using an `/llms.txt` file to provide information to help agents use a website.

Red mundial de reputación

Crowdsec

Listas de empresas y redes.



Generar conciencia

Autenticar humanos

¡Muchas Gracias!

Contacto y Enlaces de Interés



Horacio Degiorgi

horacio.degiorgi@uncuyo.edu.ar



Sistema Integrado de Documentación
bibliotecas.uncuyo.edu.ar



LA Referencia
lareferencia.info