WARC vs WACZ

La guía definitiva de formatos de preservación web



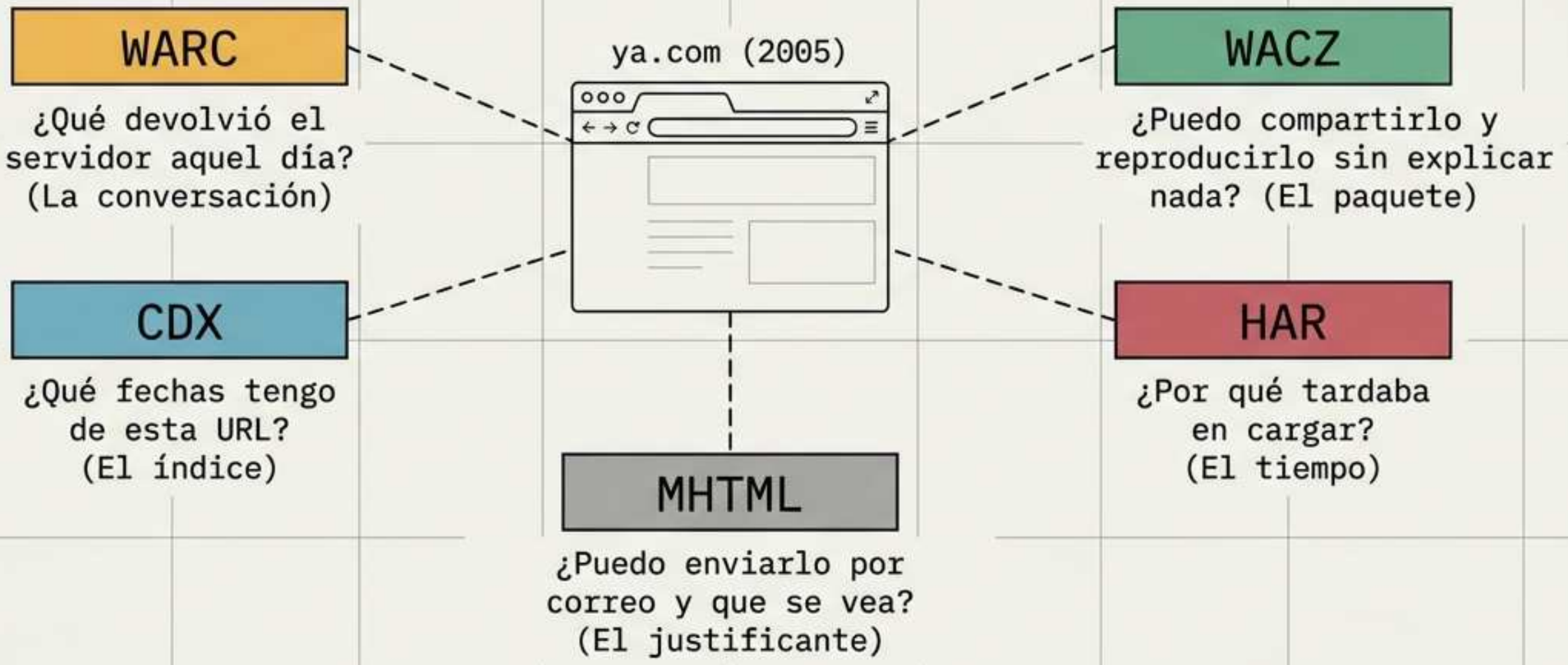
> Ejecutando análisis de ecosistema HTTP...

Basado en el análisis de Cristian Palau

Gemini Notebook

Ninguna caja contiene ‘la web’

“Elegir un formato es elegir qué preguntas no vas a poder contestar dentro de diez años.”



Gemini Notebook

Guardar un HTML no es preservar

La Ilusión (Ctrl+S)

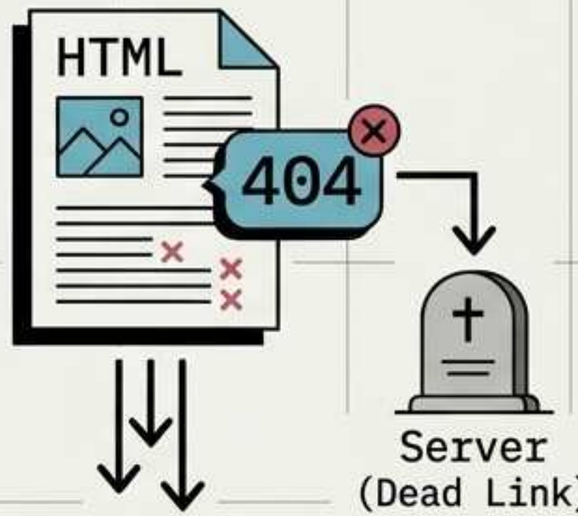


Foto parcial. Enlaces apuntan a la web viva, JS pide datos inexiistentes. Recursos perdidos.

La Realidad (Archivo)



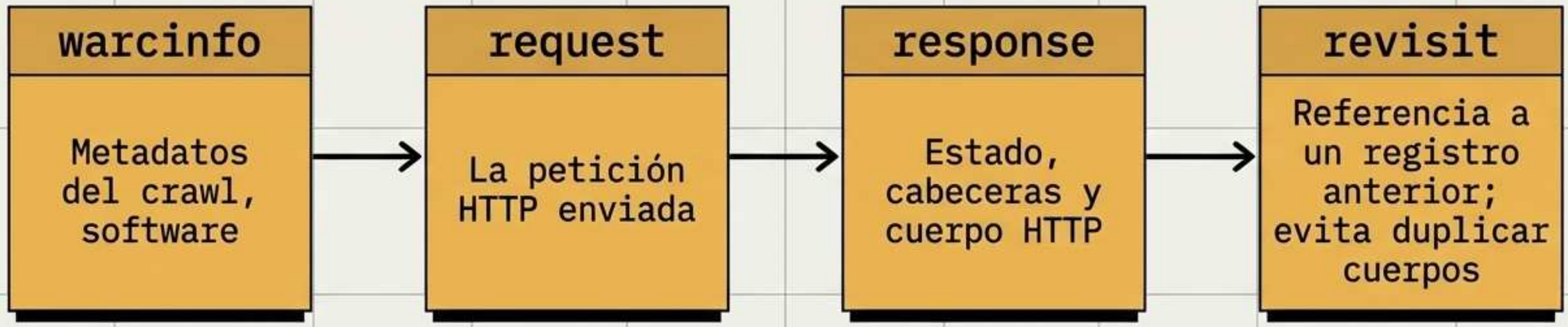
Captura de la conversación HTTP bidireccional completa. Registro inmutable.

Preservar un juego es ejecutarlo;
preservar una web es registrar sus respuestas.

Gemini Notebook

WARC: La Copia Maestra (ISO 28500)

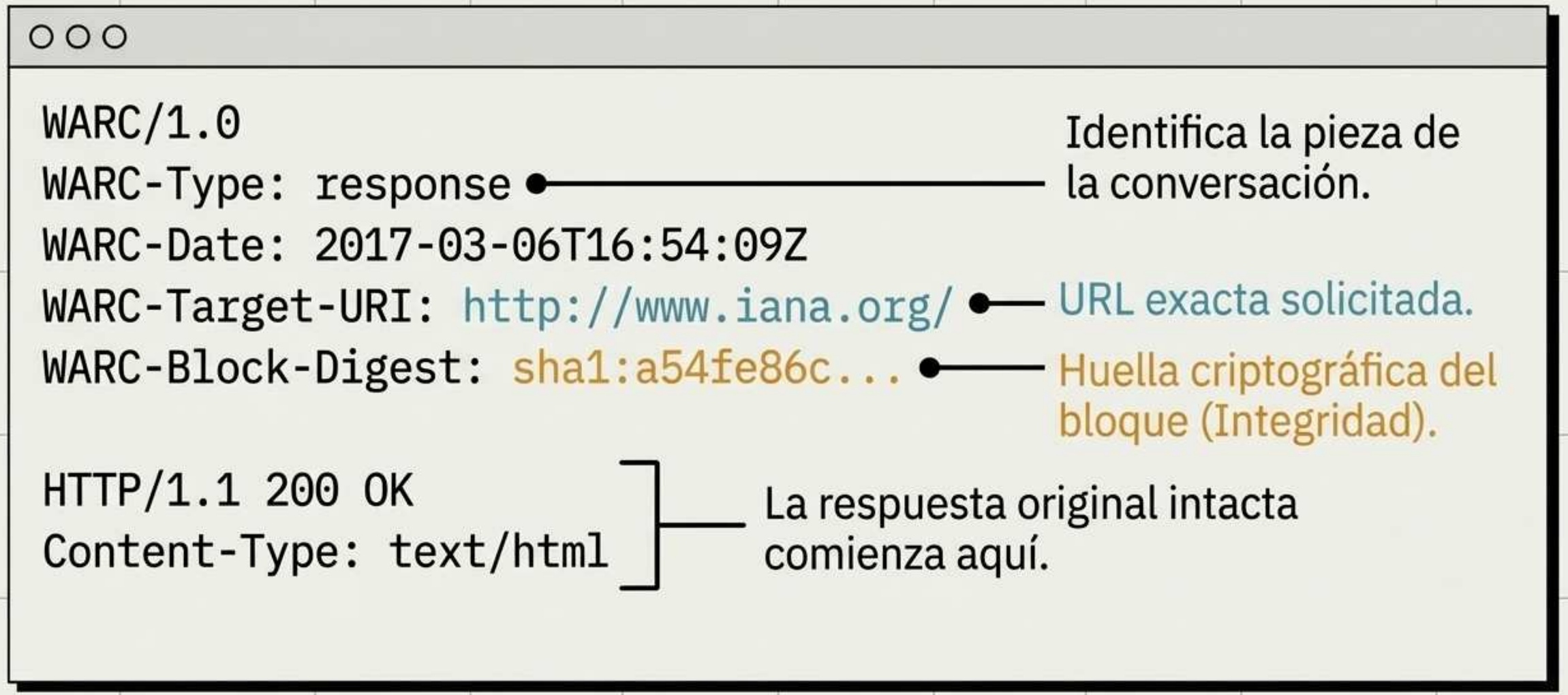
Un fichero de registros encadenados, no una carpeta comprimida.



El WARC conserva el orden y la integridad de la conversación HTTP.
Es la única prueba inmutable.

Gemini Notebook

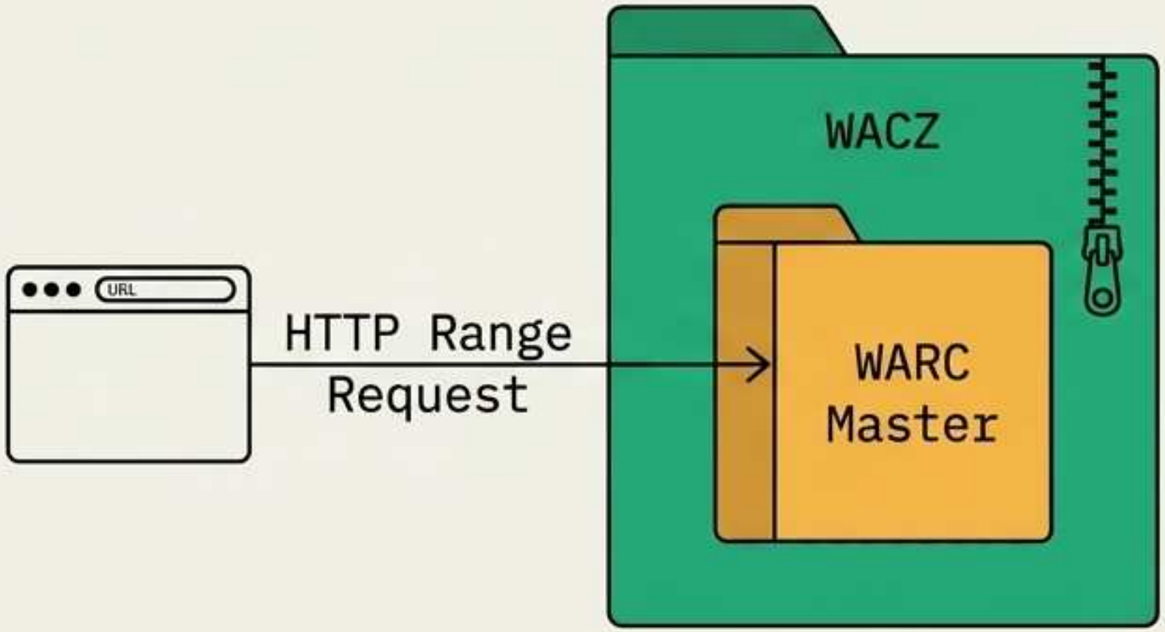
Disección de un Registro WARC



Gemini Notebook

WACZ: Un ZIP que sabe volver a la web

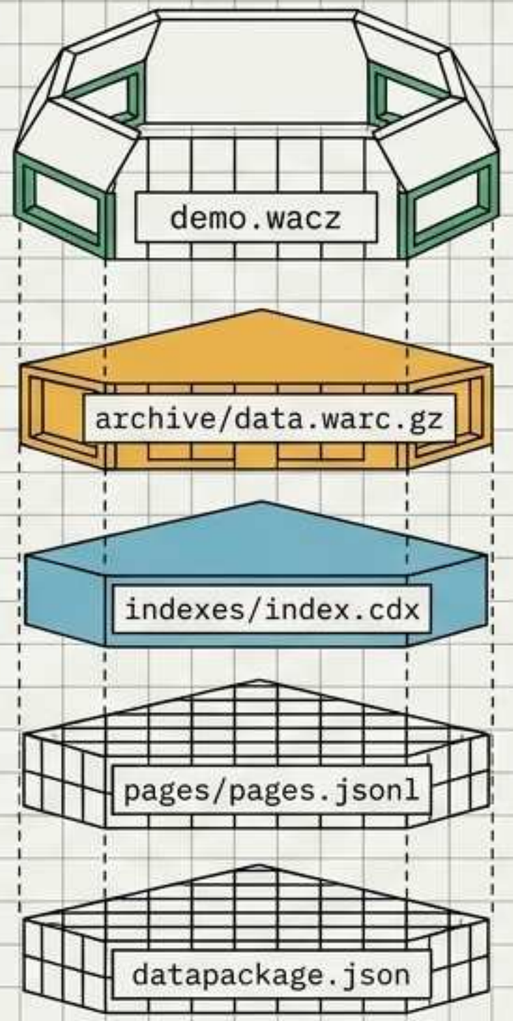
El WARC es la captura; el WACZ es la distribución y el replay.



- **No es un formato de captura:** Empaqueta un WARC existente para hacerlo cómodo de consumir.
- **Carga diferida:** Gracias a la compresión STORE y las peticiones HTTP Range, reproduce partes del archivo sin descargar el ZIP entero.
- **Fácil de compartir:** Se puede abrir directamente en **ReplayWeb.page** sin configuración de un servidor backend.

Gemini Notebook

El Despiece del WACZ



El Motor: El contenido real. Sin él, no hay nada que reproducir.

El GPS: Localiza URLs sin necesidad de leer todo el WARC.

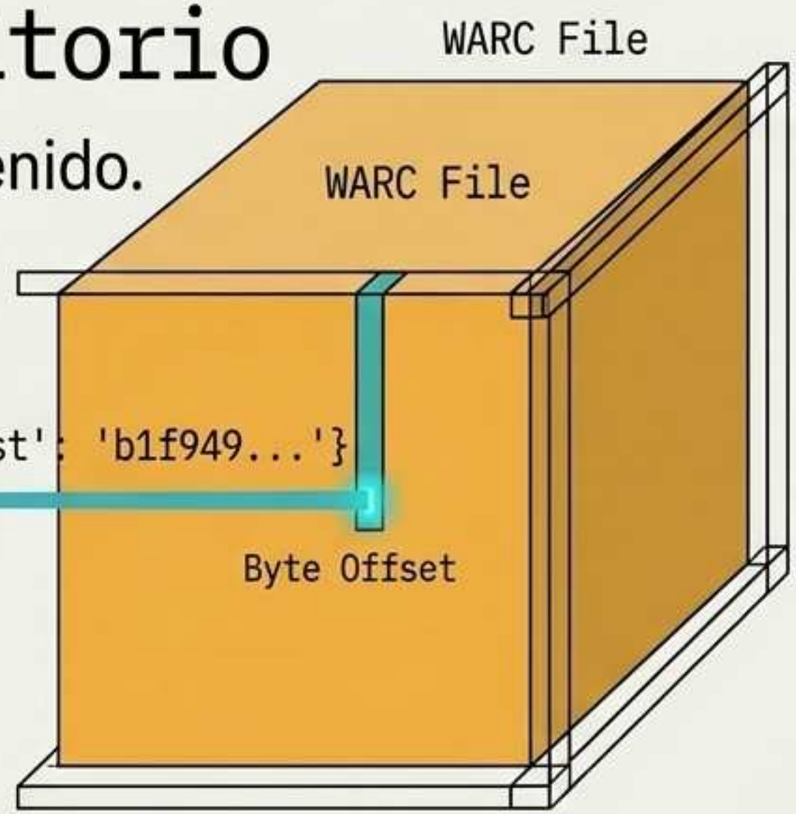
El Menú: Lista de páginas detectables para la interfaz de usuario.

El Manifiesto: Declara el perfil, los recursos, tamaños y los hashes para validación.

Gemini Notebook

CDX / CDXJ: El Mapa Sin Territorio

Te dice que hay una captura, pero no contiene el contenido.



org,iana)/ 20170306165409 response {'url': '...', 'status': '200', 'digest': 'b1f949...'}
Mechanical emitter

SURT: Dominio al revés para agrupar familias (www.example.com = example.com).

```
org,iana)/ 20170306165409 response {'url': '...', 'status': '200', 'digest': 'b1f949...'}
```

Fecha UTC de captura.

SHA-1 del cuerpo. Permite saber si dos capturas son idénticas sin abrir el WARC.

Si guardas un WARC suelto, guarda siempre su índice.

Gemini Notebook

HAR: La Traza de Diagnóstico (No un archivo)



- ✓ Perfecto para rendimiento. Incluye tiempos exactos de red.
- ✓ Formato JSON universal desde las DevTools del navegador.

- ✗ Peligro de privacidad: Filtra cookies, tokens y cabeceras de Authorization.
- ✗ No está pensado para colecciones ni para conservar relaciones entre páginas.

Veredicto: Úsalo para responder "¿por qué tardaba?", no "¿qué había en esta web?". Limpia siempre las credenciales.

Gemini Notebook

El Cementerio y Los Casos Aislados

Deprecation: **ARC (Archive)**	Deprecaton: **MHTML & .webarchive**	Deprecation: **WBN (Web Bundles)**
El abuelo de los 90 inventado por Internet Archive. Menos metadatos que WARC. No tiene tipo revisit (duplica datos).	Un sobre MIME multiparte. Todos los recursos en Base64. Crece muchísimo. Sin índice. No soporta colecciones.	Experimento abandonado de Chrome para empaquetar 'la web' offline.
Acción: Convertir. Usa `warcio recompress` para pasarlo a WARC.	**Acción:** Justificante. Útil solo para guardar una página suelta y enviarla por correo.	**Acción:** Ignorar. `ReplayWeb.page` no lo admite. Sin soporte actual.

Gemini Notebook

¿Quién archiva el mundo?

El acceso cambia, el formato backend no.

Internet Archive	Common Crawl	UK Web Archive	BNE (España)
Miles de millones de sitios. Acceso vía web.archive.org (Wayback).	Crawls mensuales abiertos. Acceso vía Descarga de WARC/Índice HTTP.	Depósito legal. Acceso vía **pywb** u OpenWayback.	Depósito legal nacional. Acceso en sedes físicas/web.

WARC (Backend Inmutable)

Insight: Si tu sitio entra en una biblioteca nacional, acabará en un WARC, no en un WACZ. WACZ es para herramientas de captura de usuario.

Gemini Notebook

Workflows de Conversión (Cheat Sheet)

Operaciones seguras usando la herramienta uv (sin instalación global).

1. ARC → WARC (Modernizar legado)

```
$ uv run --with warcio warcio recompress old.arc old.warc.gz
```

Nota: No reescribe contenido, añade los campos que ARC no tenía.

2. WARC → WACZ (Empaquetar para compartir)

```
$ uv run --with wacz wacz create -o demo.wacz example.warc
```

Nota: Mantiene el WARC intacto; añade índice CDXJ y manifiesto JSON.

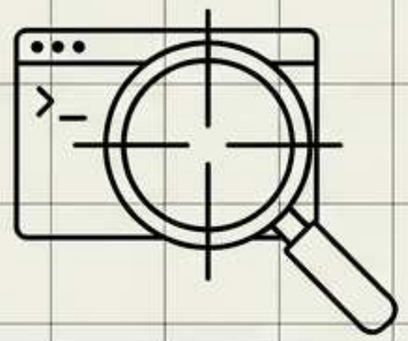
3. WARC → CDXJ (Extraer el índice)

```
$ uv run --with pywb cdx-indexer -j example.warc > example.cdxj
```

Gemini Notebook

Inspección y Reproducción Segura

Inspeccionar sin ejecutar



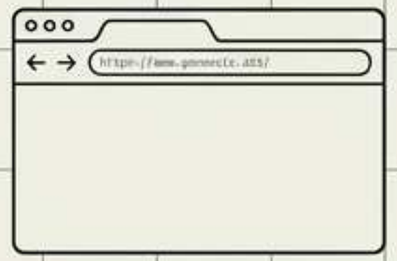
file example.warc -> Confirma el tipo MIME real.

wacz validate -f paquete.wacz manifiesto cuadran.

Verifica que los hashes del ¡Si falla, es solo un ZIP roto!

unzip -l paquete.wacz -> Lista el contenido sin extraer nada.

Reproducción



ReplayWeb.page: Funciona 100% en el navegador (cliente local). Soporta WARC, WACZ, HAR, CDX. No soporta ARC.

pywb: Servidor propio (backend) con calendario y soporte CDX avanzado.

Regla de oro: Nunca abras un HTML archivado antiguo directamente en tu navegador de trabajo.

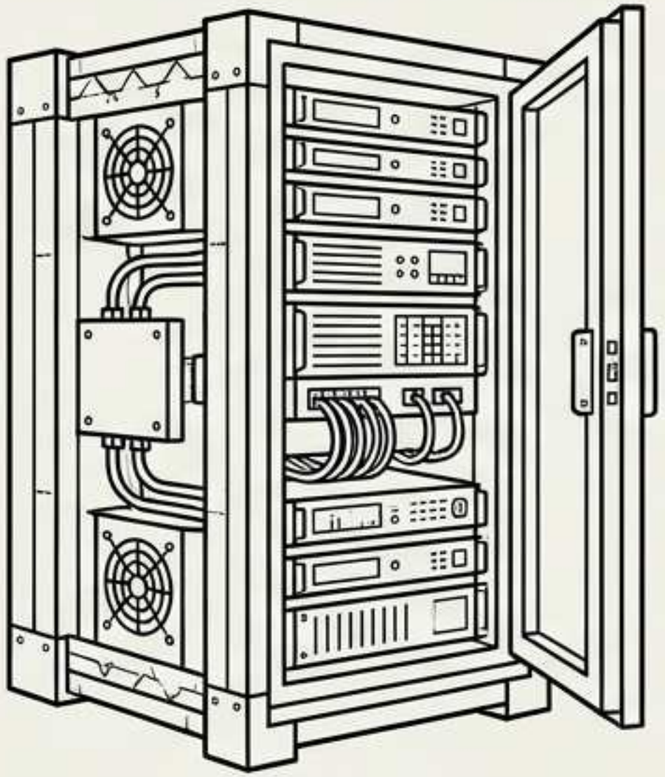
Gemini Notebook

La Matriz de Decisión Definitiva

Si necesitas...	Usa...	Por qué
Conservar a largo plazo (Copia Maestra)	[WARC]	Estándar ISO, metadatos, verificable.
Compartir y reproducir sin fricción	[WACZ]	Contiene WARC + índice + interfaz lista.
Buscar una URL por fecha	[CDXJ]	Índice ultra ligero, actúa como puntero.
Depurar tiempos de carga	[HAR]	Tiempos de red detallados, pero cuidado con PII.
Enviar un justificante rápido	[MHTML]	Un solo fichero legible en Chromium.

Gemini Notebook

El Checklist de Preservación (Regla 3-2-1)



- ☒ **Arquitectura Dual:** WARC como copia maestra segura; WACZ como copia de acceso distribuible.
- ☒ **Integridad Externa:** Hashes comprobados y guardados fuera del paquete.
- ☒ **Regla 3-2-1:** 3 copias, 2 soportes distintos, 1 fuera de línea. (Un WARC en el mismo disco que el servidor no es una copia).
- ☒ **Documentación del Contexto:** Registra la URL, fecha, crawler usado, alcance y robots.

"Dentro de diez años, el fichero seguirá ahí; las personas que sabían qué era, no. Documenta el contexto."

Gemini Notebook