Indelli Anonimizzatore

Specifiche Tecniche del Modulo:
"Anonimizzatore Atti PA"

Elaborazione Documentale, Oscuramento Selettivo XML e Conversione Headless

1. Introduzione e Obiettivi

Il modulo "Indelli Anonimizzatore Atti PA" è un applicativo web server-side progettato per automatizzare la redazione di atti amministrativi conformemente agli obblighi normativi in materia di protezione dei dati personali (GDPR) e trasparenza amministrativa (es. pubblicazione all'Albo Pretorio o in Amministrazione Trasparente).

L'obiettivo primario è consentire agli operatori della Pubblica Amministrazione di elaborare un documento in formato Word (.docx), oscurare selettivamente i dati sensibili mediante tag testuali e convertire i file risultanti in formati standard ISO (.pdf) pronti per la firma digitale, senza esporre informazioni riservate.

2. Architettura e Sicurezza degli Accessi (Session Guard)

L'applicativo non è un endpoint aperto, ma si inserisce rigidamente all'interno dell'architettura Zero-Trust dell'infrastruttura Indelli Bridge TOTP :

- - Controllo di Sessione Obbligatorio: In apertura di script, viene verificata l'esistenza delle variabili di sessione $_SESSION["auth_user"] e $_SESSION["auth_key"].
- - Rifiuto Incondizionato: Qualsiasi tentativo di accesso diretto senza autenticazione comporta l'interruzione immediata dello script e la restituzione di un errore descrittivo (HTTP Forbidden).
- - Isolamento dei File Temporanei: I file caricati e generati non risiedono in cartelle pubbliche web, ma vengono elaborati all'interno di una directory di sistema protetta (sys_get_temp_dir() / anonimizzatore_pa/).

3. Logica di Funzionamento e Algoritmi di Elaborazione

3.1 Gestione e Validazione dell'Upload

Il sistema accetta esclusivamente file con estensione .docx (MIME type OpenXML). Il nome del file originale viene filtrato tramite espressioni regolari (preg_replace) per rimuovere caratteri speciali o potenziali vettori di path traversal, associandolo a un timestamp univoco per evitare collisioni.

3.2 Motore di Redazione OpenXML (Omissis)

I file .docx moderni sono archivi ZIP contenenti file XML strutturati. L'algoritmo di anonimizzazione sfrutta questa peculiarità:

- 1. Estrazione Strutturale: Tramite la libreria nativa PHP ZipArchive, viene estratto il file principale del testo (word/document.xml).
- 2. Pattern Matching (Regex UTF-8): Utilizza un'espressione regolare multibyte orientata alla ricerca di marcatori racchiusi tra doppie parentesi quadre: /\[\[(.*?)\]\]/u (es. [[Mario Rossi]]).
- 3. Sostituzione Polimorfica: Ogni corrispondenza trovata viene rimpiazzata programmaticamente con la stringa fissa 'OMISSIS'.
- 4. Re-incapsulamento: Il file XML modificato viene reinserito nell'archivio ZIP, generando in parallelo il file Originale e il file Omissis.

3.3 Pipeline di Conversione Headless (DOCX to PDF)

Per soddisfare gli standard di immodificabilità e preparazione alla firma digitale, il sistema integra una funzione di conversione basata su LibreOffice in modalità Headless:

libreoffice --headless --convert-to pdf [file.docx] --outdir [dir]

La funzione verifica l'effettivo codice di uscita del processo ($returnVar === 0) e la presenza fisica del file PDF generato, gestendo eventuali eccezioni operative sul server.

4. Gestione I/O e Sicurezza dei Download

Il trasferimento dei file generati verso il client avviene tramite un sistema di streaming controllato via parametri GET:

- - Headers di Sicurezza: Vengono iniettati header HTTP mirati (Content-Type, Content-Length, Cache-Control: must-revalidate) per forzare il download diretto ed evitare il caching non autorizzato.
- - Ridenominazione Istituzionale: I file scaricati vengono automaticamente rinominati in modo standardizzato (es. ATTO_OMISSIS.pdf o ATTO_ORIGINALE_FIRMABILE.pdf).
- - State Management: Ad ogni nuovo caricamento o conversione, le variabili di sessione vengono aggiornate o ripulite, evitando sovrapposizioni di dati.

5. Matrice di Sicurezza del Modulo

Vettore di Minaccia / Rischio	Soluzione Implementata	Livello di Efficacia
Accesso Non Autenticato alle API	Controllo rigoroso della sessione Zero-Trust ($_SESSION["auth_user"]).	Totale (Blocco a monte)
Iniezione di file malevoli (Web Shell)	Restrizione ferrea dell'estensione al solo .docx e sanitizzazione tramite regex.	Alta
Esposizione pubblica dei dati originali	Archiviazione dei file temporanei in directory di sistema (sys_get_temp_dir) non esposte nel web root.	Alta
Errori di omissis manuali (Dimenticanze)	Automazione totale basata su tag espliciti ([[...]]) tramite manipolazione XML nativa.	Alta