DetectaNIF: Manual de uso

# Acerca de

DetectaNIF es una utilidad para localizar ficheros candidatos a contener datos de carácter personal. Para ello, busca la presencia de cadenas de caracteres coincidentes con el formato de NIF/NIE de persona física.

# Modo de funcionamiento

*DetectaNIF* busca recursivamente en todas las carpetas que se le hayan indicado. Para cada fichero existente en cada carpeta determina:

* Si tiene alguna de las extensiones de fichero configuradas para ser elegible.
* Si tiene el tamaño mínimo configurado para ser elegible.
* Si el tiempo transcurrido desde su creación supera el mínimo configurado para ser elegible.

En caso afirmativo, el fichero es incluido en una cola para su exploración. La cola de ficheros es atendida por un número de hilos concurrentes (configurable). Cada hilo toma un fichero, explora su contenido y escribe el resultado en el fichero de salida con formato CSV (codificado en UTF-8). La exploración tiene estas características:

* Es binaria (no necesita aplicaciones externas para abrir cada fichero).
* Busca cadenas coincidentes con el formato de NIF/NIE con codificación UTF-8 o UTF-16.
* No se exploran los últimos 17 bytes del fichero, puesto que no pueden contener una cadena de NIF/NIE codificada en UTF-16, pero podría contener una cadena NIF/NIE en UTF-8. Por tanto, hay un margen de error de un NIF.
* En ficheros comprimidos es posible detectar las cadenas NIF/NIE, pero su número será inferior al real.
* No arroja resultados en ficheros cifrados, si bien, no está afectado por el cifrado de volumen (tipo “BitLocker”).

La exploración tiene dos límites (configurables) con el propósito de disminuir su tiempo de ejecución. En primer lugar, se finaliza la exploración cuando ya se ha alcanzado un cierto número de cadenas NIF/NIE (parámetro “MaxNIF”). Por otra parte, se finaliza la exploración si la primera porción del fichero no contiene ninguna cadena NIF/NIE:

* La porción en concreto se calcula como un porcentaje del tamaño del fichero, configurado en el parámetro “MaxSinNIF”.
* No obstante, siempre se procesan bloques enteros de 64 kilobytes.

## Rutas de fichero

*DetectaNIF* es capaz de manejar rutas largas de fichero (hasta 65535 bytes) y en formato UNICODE (UTF-8/UTF-16). Son válidas las rutas UNC (“\\servidor\recurso”). Se recomienda usar rutas absolutas.

# Fichero de configuración

Todos los parámetros de exploración se suministran en un fichero de configuración con formato CSV sin necesidad de cabeceras. El carácter separador de campos es el punto y coma (“;”). Se recomienda encerrar entre comillas dobles (“) todos los parámetros consistentes en rutas de fichero o carpeta. Los parámetros no reconocidos son ignorados, por tanto, se pueden introducir líneas a modo de comentario siempre que respeten el formato CSV. Por ejemplo:

FICHERO DE CONFIGURACIÓN DE EJEMPLO;

GenerarResumen;true;

IncluirErrores;true;

NumeroHilos;4;

MinDiasCreado;365;

MinBytesTamaño;3096;

MaxNIF;1000;

MaxSinNIF;50;

Extensiones;"xls xlsx accdb mdb csv txt xlsxm xlsxb mdx accda mde accde ade db dbf prn dif tab adp";

Carpeta;”L:\recursos”;

Carpeta;”E:\recursos”;

Salida; “C:\temp\candidatos NIF.csv”;

Puede generar un fichero de configuración ejecutando:

detectaNIF –plantilla cfg.csv

A continuación, debe editar el fichero (“cfg.csv” en el ejemplo). La utilidad no avisa de la existencia de parámetros no reconocidos: asegúrese de que están correctamente escritos.

## Parámetro “GenerarResumen”

Indica si se desea generar un fichero adicional a modo de resumen de la exploración (recomendado). El fichero tendrá el mismo nombre y ubicación que el fichero de salida (parámetro “Salida”) pero con extensión “.log”.

Valores admitidos: “true” o “false” (valor por omisión).

No es obligatorio. Si se repite, solamente se tiene en cuenta el último valor.

## Parámetro “IncluirErrores”

Indica si se desea dejar constancia, en el fichero de salida, de los ficheros que no pudieron ser explorados a causa de cualquier error de acceso al fichero. Generalmente, estos errores se deben a insuficiencia de permisos.

Valores admitidos: “true” (valor por omisión) o “false”.

No es obligatorio. Si se repite, solamente se tiene en cuenta el último valor.

## Parámetro “NumeroHilos”

Indica el número de exploraciones concurrentes.

Valores admitidos: número entero entre uno (valor por omisión) y diez, ambos inclusive.

Se recomienda hacer coincidir este parámetro con el número de núcleos físicos del microprocesador. No es obligatorio. Si se repite, solamente se tiene en cuenta el último valor.

## Parámetro “MinDiasCreado”

Indica el número mínimo de días que han de transcurrir desde la fecha de creación de un fichero para que éste sea elegible para su exploración.

Valores admitidos: entero mayor o igual que cero (valor por omisión).

No es obligatorio. Si se repite, solamente se tiene en cuenta el último valor.

Se recomienda fijarlo en 365 días.

## Parámetro “MinBytesTamaño”

Indica el tamaño mínimo que ha de tener un fichero para que sea elegible para su exploración.

Valores admitidos: entero mayor o igual que cero (valor por omisión).

No es obligatorio. Si se repite, solamente se tiene en cuenta el último valor.

## Parámetro “MaxNIF”

Indica el número máximo de cadenas coincidentes con el formato de NIF/NIE que serán contabilizadas. Superado este límite, se finalizará la exploración del fichero.

Valores admitidos: entero mayor o igual que veinte. Por omisión, infinito.

No es obligatorio. Si se repite, solamente se tiene en cuenta el último valor.

Se recomienda fijar este parámetro en 1.000.

## Parámetro “MaxSinNIF”

Indica, en porcentaje, la porción inicial del fichero que debe contener alguna cadena NIF/NIE para continuar con la exploración.

Valores admitidos: entero entre 40 y 100 (valor por omisión), ambos inclusive. Los valores admitidos evitan falsos negativos.

No es obligatorio. Si se repite, solamente se tiene en cuenta el último valor.

Se recomienda fijar este parámetro en 50.

## Parámetro “Extensiones”

Indica las extensiones que puede tener un fichero para que sea elegido para su exploración.

Todas las extensiones deben indicarse en una sola línea, separadas por un espacio y sin punto. Se admiten los caracteres comodines (asterisco e interrogación). Por ejemplo, para explorar todos los ficheros, debe indicarse:

Extensiones;\*;

No obligatorio (por omisión, se exploran todos los ficheros). Si se repite, solamente se tiene en cuenta el último valor. Se recomienda fijar este parámetro en las extensiones de ficheros ofimáticos:

Extensiones;xls xlsx accdb mdb csv txt xlsxm xlsxb mdx accda mde accde ade db dbf prn dif tab adp;

## Parámetro “Salida”

Indica el nombre y ruta del fichero a generar con los resultados. El fichero tendrá formato CSV codificado en UTF-8. Obligatorio, aunque también puede ser proporcionado en la línea de comandos.

## Parámetro “Carpeta”

Indica la ruta a una carpeta a explorar (recursivamente).

Si se repite, se tienen en cuenta todos los valores. Por tanto, es posible configurar varias carpetas.

Alguna de las carpetas configuradas puede ser ignorada si:

* No es una ruta bien formada a una carpeta
* Se ha configurado otra carpeta que la contiene
* La carpeta no existe.

# Línea de comandos

Para ejecutar la exploración, indique el fichero de configuración. Por ejemplo:

detectaNIF –cfg “C:\temp\detectaNIF.cfg”

Además, es posible indicar el fichero de salida:

detectaNIF –cfg “C:\temp\detectaNIF.cfg” –salida “C:\temp\otrasalida.csv”