SEPINF:IPED

De wiki

Índice

Indexador e Processador de Evidências Digitais

Descrição

Programa linha de comando em java originalmente desenvolvido para indexar relatórios do FTK 1.8 (convertidos pelo AsAP3) e relatórios do FTK 3+.

Atualmente apresenta diversas funcionalidades presentes em softwares forenses comerciais, servindo como alternativa eficiente e de código aberto na maioria dos casos.

Disponível no WebDAV, na pasta Indexador e Processador de Evidências Digitais (https://sepinf.ditec.dpf.gov.br/dav/Softwares/Analise%20de%20Midias/Indexador%20e%20Processador%20de%20Evidências%20Digitais/).

Objetivos

Facilitar a visualização de Laudos de extração de dados por parte da equipe de investigação e melhorar a qualidade de Laudos contendo resultados de pesquisas por palavras-chave.

Facilitar a análise dos dados em soluções de SARD, como a descrita em Triagem de Mídias SETEC/DF.

Ser uma alternativa aos softwares comerciais forenses na maioria dos casos periciais, por exemplo, em casos que necessitem de função de busca por palavras-chave ou recuperação de dados apagados.

Configuração

O principal arquivo de configuração da ferramenta é o IPEDConfig.txt. Lá pode ser configurado o cálculo de hash, indexação do conteúdo dos arquivos (indexFileContents), indexação do espaço não alocado (indexUnallocated), cálculo de assinatura (processFileSignatures), carving (enableCarving), expansão de containers (expandContainers), dentre diversas outras opções comentadas no próprio arquivo. A seguir são detalhadas algumas configurações importantes.

Outro arquivo importante de configuração é o LocalConfig.txt onde são definidas configurações específicas do ambiente, como diretório temporário indexTemp, se o diretório temporário está em SSD, número de workers de processamento, e caminhos para as bases de hashes do IPED (NSRL do NIST), de pornografia infantil do LED, base de detecção de nudez do LED, e caminho para o TskDatamodel.jar compilado em sistemas Linux.

Relatórios do FTK 3+

Função específica para indexar relatórios do FTK para facilitar a revisão de relatórios de extração de dados pelo solicitante do exame.

- É necessário gerar o relatório utilizando o ID dos arquivos como nomenclatura e o IPED obtém as propriedades originais dos itens do banco de dados usando seus IDs (caso seja de interesse, na busca o usuário pode exportar todos os arquivos com seus nomes originais).

Também é preciso editar o arquivo conf/FTKDatabaseConfig.txt ou fornecer os seguintes parâmetros na tela de preferências do indexador no AsAP 4.4+:

- Banco de dados: Oracle, Postgres ou SQLServer
- IP e porta do servidor de banco de dados
- Nome da base de dados: FTK2 ou ADG
- Versão do FTK: automático a partir do Indexador 2.9
- Usuário e senha do banco de dados

Observe que este usuário é diferente do que você utiliza no FTK, portanto a senha provavelmente será diferente. No Oracle, para verificar se você sabe a senha, execute "sqlplus" no prompt de comando e faça login como "sys as sysdba", no caso do Oracle.

No caso de a senha do usuário sys não ser conhecida, pode ser renomeado o arquivo de senhas C:\Oracle\Product\10.2.0\AccessDataDB\database\PWDftk2.ora e gerado outro de mesmo nome com o utilitário orapwd, fornecendo a nova senha.

```
orapwd file=C:\Oracle\Product\10.2.0\AccessDataDB\database\PWDftk2.ora password=senha
```

Processamento de Imagens Forenses

O software é capaz de acessar o conteúdo de dispositivos físicos e imagens forenses DD, 001, E01 e ISO por meio da suite forense The Sleuthkit (TSK) e Libewf, os quais são utilizados para acessar o conteúdo das imagens e decodificação dos sistemas de arquivos. Além disso, por meio do Sleuthkit, é realizada automaticamente recuperação simples de arquivos apagados das tabelas de arquivos dos sistemas de arquivos. Por meio do TSK também é acessado o espaço não alocado, o qual é indexado e submetido a data carving otimizado pelas tarefas de processamento específicas do IPED.

É incluída uma versão do Sleuthkit para Windows com patches relativos a correções de bugs e otimizações. Para uso em sistemas Linux, recomenda-se a aplicação dos patches incluídos na pasta sources na raiz do IPED.

Categorização

A categorização dos arquivos é realizada principalmente via análise de assinatura pela biblioteca Apache Tika (http://tika.apache.org). A biblioteca retorna o mimeType do arquivo, o qual é mapeado para uma categoria conforme configurado no arquivo conf/CategoriesByTypeConfig.txt.

Caso deseje definir um novo tipo (mimeType) de arquivo por assinatura, nome ou extensão, adicione a definição no arquivo conf/CustomSignatures.xml.

Além disso, a categoria dos arquivos pode ser refinada com base em qualquer propriedade, como caminho, tamanho, datas, deletado, etc. Para isso, utilize o arquivo conf/CategoriesByPropsConfig.txt, o qual utiliza linguagem javascript para permitir flexibilidade nas definições.

Detecção de Arquivos Criptografados

A partir da versão 3.3 é realizada automaticamente detecção de arquivos criptografados dos seguintes tipos: pdf, office97 (doc, xls, ppt), office2007 (docx, xlsx, pptx), openoffice (odt, ods, odp), zip, rar, 7z e pst. Os arquivos identificados como cifrados podem ser acessados por um filtro preconfigurado.

A partir da versão 3.13, é realizada detecção de itens cifrados por entropia. É utilizado o algoritmo LZ4, extremamente rápido, para comprimir os itens e a taxa de compressão é avaliada. Entretanto isso gera alguns falsos positivos, pois esse algoritmo não comprime tanto. Poderia ser usado o algoritmo Deflate (ZIP) que comprime melhor, porém haveria impacto na performance de processamento. Para diminuir a quantidade de falsos positivos, são considerados apenas itens maiores que 100MB, com erro de parsing ou sem parser específico.

Expansão de Containers

Para a expansão de containers, é utilizada a biblioteca Apache Tika (http://tika.apache.org), que fornece suporte para zip, tar, ar, arj, jar, gzip, bzip, bzip2, xz, 7z, z, cpio, dump, formatos office, rtf e pdf. Caso se queira extrair imagens embutidas em PDFs, é necessário habilitar a opção processImagesInPDFs em conf/AdvancedConfig.txt. Além disso, a biblioteca foi estendida e foram implementados parsers para dbx, pst, ost, mbox, eml, rar, iso e edb. Entretanto, atualmente a implementação tem a limitação de não recuperar emails apagados de dentro de dbx.

Para habilitar a expansão de containers, habilite o parâmetro expandContainers no arquivo IPEDConfig.txt. As categorias a serem expandidas podem ser alteradas no arquivo conf/CategoriesToExpand.txt. A expansão extrai os subitens para a pasta "subitens" e é recursiva, podendo utilizar espaço considerável. O hash dos arquivos é usado como nomenclatura para economizar espaço.

Recomenda-se configurar uma pasta de saída (-o) do processamento num disco diferente daquele que contém as imagens/dados sendo processados, para minimizar acessos concorrentes de leitura dos dados e escrita dos subitens expandidos.

Cálculo de múltiplos Hashes

Atualmente o software suporta os seguintes algoritmos de hashes criptográficos: md5, sha-1, sha-256, sha-512, edonkey.

Consulta a Base de Hashes (KFF)

Na versão 3.3 foi incluída função de consulta a base de hashes local para alertar ou ignorar arquivos. Podem ser importadas bases no formato NSRL (parâmetro -importkff), a qual é armazenada em formato pré-indexado para consultas. É altamente recomendado configurar a base num disco SSD, sob pena de degradar o tempo de processamento. É necessário configurar o caminho da base (opção kffDb) no arquivo de configuração. Os arquivos encontrados na base recebem um atributo kffstatus com valor "alert" ou "ignore", sendo que os ignoráveis podem ser excluídos do caso se habilitado (excludeKffIgnorable). É possível alterar a lista de programas cujos arquivos devem receber o status de alerta no arquivo conf/KFFTaskConfig.txt.

Também há uma função específica para consultar hashes na base de pornografia infantil no LED, bastando configurar o caminho da base no arquivo de configuração principal (ledWkffPath). A vantagem é que a base pode ser atualizada facilmente bastando apontar para uma versão nova do LED, sem necessidade de importação. Os arquivos encontrados nessa base são adicionados a categoria de alerta específica.

Obs: o parâmetro --importKff é utilizado apenas para importar bases no formato original do NSRL pela primeira vez. A base disponível no DAV (na pasta do IPED) já está importada/codificada no formato do IPED, bastando habilitar a opção “enableKff” no arquivo IPEDConfig.txt e informar o caminho da base na opção “kffDb” do arquivo LocalConfig.txt.

Indexação

Antes da indexação com a biblioteca Apache Lucene, é realizada a extração de texto dos arquivos com a biblioteca Apache Tika (http://tika.apache.org). Dentre os formatos suportados, podem ser citados: MS Office (doc, docx, xls, xlsx, ppt, pptx e similares), OpenOffice (odt, ods, etc), Apple iWork (key, pages, numbers), PDF, HTML e XML, RTF e TXT, e-mails (RFC822 e Outlook MSG) e metadados de audio (midi, mp3), imagens (bmp, jpg, psd, png, tif, etc) e vídeos (flv, mp4 e derivados e ogg e derivados), dentre outros.

Além disso, foram criados parsers adicionais para MS Access, xBase, SQLite, Registro do Windows, atalhos LNK (PCF Gabriel), Known.met e ShareL/H.dat (PCF Wladimir), Library1/2.dat e WhatsApp (PCF Pfeifer), Skype (PCF Patrick), bancos EDB, históricos Index.dat, além de um extrator de strings brutas ISO-8859-1, UTF-8 e UTF-16 utilizado como fallBackParser com todos os demais tipos de arquivo não suportados pelo TIKA, como binários, desconhecidos, corrompidos e espaço não alocado.

É possível habilitar/desabilitar parsers no arquivo conf/ParserConfig.xml, desabilitar a indexação de binários e desconhecidos (indexUnknownFiles) ou desligar a indexação do espaço não alocado (indexUnallocated), o que pode ser interessante dependendo do volume de dados (como em casos de triagem de dados - SARD), pois deixa o processamento mais rápido, resulta num índice muito menor e não apresenta hits de difícil interpretação em arquivos como pagefile, system restore, espaço não alocado, etc.

A partir da versão 3.3, foi incluído teste de aleatoriedade (entropia) antes de indexar trechos de arquivos desconhecidos ou não alocado, o que melhora bastante a eficiência de indexação desses arquivos. Entretanto, eventualmente podem ser perdidos hits cercados por conteúdo "aleatório".

OCR

O OCR utiliza o programa Tesseract 3.02 e é executado sobre imagens (jpg, tif, png e bmp) e arquivos PDF com pouco texto, fazendo parte da tarefa de parsing dos arquivos. É o processamento mais pesado e utiliza bastante a CPU, podendo demorar alguns segundos por imagem e retardar em até 10x o tempo de processamento. Por isso fica desabilitado por padrão e pode ser habilitado no arquivo IPEDConfig.txt (enableOCR).

Os resultados podem variar bastante, dependendo da qualidade e resolução das imagens, tamanho e tipo das fontes utilizadas.

O número de caracteres reconhecidos é armazenado no metadado OCRCharCount, permitindo localizar imagens contendo textos, como digitalizações, com pesquisas como ocrcharcount:[100 TO *] a qual retorna imagens com 100 ou mais caracteres reconhecidos, ou simplesmente ordenando por esse atributo.

O OCR de itens duplicados (segundo o hash) é reaproveitado, otimizando bastante o processamento em alguns casos.

Data Carving

Para ativar o carving, habilite o parâmetro enableCarving no arquivo IPEDConfig.txt. No arquivo conf/CarvingConfig.txt (enableCarving) há opções para incluir ou excluir arquivos do processamento, por exemplo, realizar carving apenas sobre o espaço não alocado e/ou sobre itens alocados (por exemplo, pagefile, thumbs.db, system restore, executáveis, desconhecidos, etc). A configuração padrão é bastante inclusiva, excluindo basicamente os containers com expansão suportada para evitar a recuperação de itens duplicados.

Atualmente estão configuradas assinaturas para os seguintes tipos de arquivo: sqlite, bmp, emf, gif, png, jpg, webp, html, pdf, ole e derivados (doc, xls, ppt, msg, thumbs.db), index.dat, rar, zip e derivados (office 2007, OpenOffice, iWork, xps, cdr), eml com anexos base64, avi, wmv, mp4, 3gp, mov, flv, mpeg (tem falsos negativos), wma, wav, midi, cda, shareL/H.dat(ares), podendo ser adicionadas novas assinaturas no arquivo conf/CarvingConfig.txt.

O algoritmo de carving utilizado é bastante eficiente e não degrada com o número de assinaturas pesquisadas, levando o mesmo tempo caso buscadas 1 ou 10000 assinaturas. É proporcional apenas ao volume de dados processado e ao número de assinaturas encontradas (e não de pesquisadas), conseguindo a atingir taxas entre 300 a 700 GB/h de carving na estação promo2, geralmente limitado pelo I/O.

KnownMetCarving e KFFCarving

A partir da versão 3.10, foram integrados o KnownMetCarving (carving para known.met do Emule) e o KFFCarving (carving de itens presentes na base do LED) pelo PCF Wladimir. A vantagem do último é que pode recuperar tipos não recuperados pelo carving tradicional (como mpeg) e mesmo itens fragmentados ou sobrescritos recuperados parcialmente, cujos hashes não dariam hit na base de pornografia infantil do LED, recebem uma nomenclatura especial (CarvedKFF), indicando que se tratam de fragmentos de arquivos presentes na base do LED.

Miniaturas de Imagens

A partir da versão 3.9 são geradas miniaturas de imagens durante o processamento por padrão, o que pode ser desabilitado. A visualização de imagens na galeria fica instantânea. Além disso, tornou-se possível filtrar imagens (e vídeos) sem miniaturas geradas (normalmente por estarem corrompidas) por meio de filtro preconfigurado.

A partir da versão 3.13, alterou-se o GraphicsMagick (GM) para o ImageMagick (IM) para visualizar centenas de formatos de imagens não suportados pelo Java. Nessa versão, o custo de processamento na geração de miniaturas foi bastante reduzido, gerando-se miniaturas de imagens parciais de carving de formatos comuns (JPG, PNG, BMP e GIF) via java puro. Pode-se utilizar o GraphicsMagick (GM) no lugar do IM, porém ele suporta metade dos formatos que o IM e é mais lento na geração de miniaturas de alguns formatos, principalmente WMF.

Esse e outros parâmetros, como tempo de timeout na geração de thumbs, tamanho dos thumbs, número de threads da galeria (importante ao desabilitar geração de thumbs durante o processamento), podem ser alterados em conf/ImageThumbsConfig.txt

Miniaturas de Vídeos

(contribuição PCF Wladimir)

Na versão 3.4 foi incluída função para extração de cenas de vídeos (enableVideoThumbs), a qual utiliza o software MPlayer. Os parâmetros da extração de cenas, como resolução, número de quadros extraídos, podem ser alterados no arquivo conf/VideoThumbsConfig.txt. As miniaturas de vídeos também são exibidas na galeria de imagens, sendo recomendado "diminuir" o número de itens exibidos na galeria para aumentar o tamanho das cenas dos vídeos, permitindo uma boa visualização.

Na versão 3.5 foi incluída a opção de se exportar apenas os thumbnails (tb funciona para imagens) de um ou mais marcadores (ou categorias, caso não haja marcadores) Na geração do relatório, utilize a opcao [[-nocontent "Nome do marcador][-nocontent "Nome do Marcador 2"]...].

Detecção de Imagens Explícitas (DIE)

(contribuição PCF Wladimir)

A partir da versão 3.9, foi integrado o módulo de detecção de nudez do LED (https://wiki.ditec.dpf.gov.br/SEPINF:LED). A execução é rápida, normalmente retardando em apenas 5% o tempo de processamento total. São criados os atributos scoreNudez (1 a 1000) e classeNudez (1 a 5), permitindo ordenação das imagens por pontuação. O classeNudez apenas utiliza um intervalo de valores menor para facilitar ordenações secundárias, pelo caminho por exemplo, o que pode ser interessante em alguns casos.

Foi criado um filtro "Imagens com Possível Nudez", que realiza um corte simplista de imagens com scoreNudez acima de 500, mas não é recomendado seu uso indiscriminado devido a falsos negativos, considere o uso da ordenação.

Nos testes o algoritmo de detecção mostrou uma ótima relação precisão x cobertura comparativamente a outros softwares forenses comerciais e de código aberto.

Detecção de Idiomas

A partir da versão 3.13, foi adicionada e fica habilitada por padrão a detecção de idioma via optimaize language-detect (71 idiomas suportados). São adicionados metadados no grupo language ao itens, que indicam os 2 idiomas mais prováveis e os scores/probabilidades de cada idioma detectado.

Expressões Regulares

A partir da versão 3.13, foi adicionada função de localização de expressões regulares durante o processamento. Por padrão, já estão configuradas algumas expressões para localização de CPF, CNPJ, PisPasep, CNH, email, URL, IP, valores monetários, contas bancárias, boletos, cartão de crédito, iban, swift, título de eleitor. Quando existente, é checado o dígito verificador para as expressões padrão. Novas expressões regulares podem ser adicionadas em conf/RegexConfig.txt do profile utilizado.

As ocorrências encontradas das expressões regulares são adicionadas ao metadado Regex:NOME_REGEX, onde NOME_REGEX é o nome configurado da expressão regular. Dessa forma, os itens podem ser filtrados na aba de metadados de acordo com as ocorrências encontradas por cada expressão regular.

As expressões regulares são pesquisadas no texto extraído após a decodificação dos arquivos. Assim são encontradas ocorrências em formatos complexos como pdf, office2007, pst, dbx, imagens (com ocr ligado), etc, o que traz resultados muito superiores em relação à outras ferramentas que buscam as regex nos dados brutos dos arquivos.

Reconhecimento de Entidades Mencionadas

A partir da versão 3.13, foi adicionada função de reconhecimento de entidades mencionadas via StanfordCoreNLP. Essa função permite identificar nomes de pessoas, organizações e lugares nos textos por meio de técnicas de processamento de linguagem natural. Ainda não há modelo de treinamento para o português, sendo utilizado o inglês para idiomas sem modelo específico, o qual surpreendentemente traz resultados bastante razoáveis.

Essa função pode aumentar o tempo de processamento em até 4x, então não deve ser habilitada de forma indiscriminada. Para habilitar, configure o parâmetro enableNamedEntityRecogniton no arquivo IPEDConfig.txt. Além disso, é necessário baixar pelo menos o modelo de treinamento para o inglês de https://stanfordnlp.github.io/CoreNLP/history.html o qual totaliza 1GB de tamanho atualmente, copiando-o para a pasta ../optional_jars para ser carregado.

As entidades mencionadas encontradas nos textos são adicionadas ao metadado NER_NomeEntidade, podendo ser melhor visualizadas na aba de agrupamento por metadados. Opções avançadas de função podem ser configuradas no arquivo conf/NamedEntityRecognitionConfig.txt do profile utilizado.

Profiles de Processamento

Na v3.12, foi incluída na linha de comando a opção -profile, para a qual pode ser passada uma pré-configuração de processamento. Por padrão são incluídos os seguintes profiles: forensic, pedo, fastmode, blind.

Caso essa opção não seja utilizada, é utilizada a configuração default de processamento, que por padrão não faz carving e nem OCR, por exemplo. Para criação de um novo profile customizado, basta copiar um dos existentes dentro da pasta "profiles" (na raiz do IPED), renomear conforme desejado e alterar as opções de processamento.

A seguir são detalhados esses profiles pré-configurados.

fastmode

Ativada via --fastmode até a versão v3.11. A partir da v3.12 é ativada via "-profile fastmode". Ela realiza um rápido processamento, normalmente em poucos minutos, podendo ser usada em locais de busca, por exemplo, ou para realizar um preview dos dados antes de um exame pericial.

Essa opção usa configurações mínimas de processamento: não calcula hash, nem assinatura, não indexa conteúdo, não faz carving, não gera miniaturas de vídeos, não expande conteineres, nem nenhuma outra tarefa que acessa o conteúdo dos arquivos. Permite assim um processamento rápido, limitado pela decodificação do FS pelo sleuthkit.

Porém, continua a recuperar itens apagados das tabelas de arquivos, a categorizar os itens (por extensão), permite ordenações e filtros por atributos (nome, extensão, tamanho, etc), permite visualizar os itens, utilizar a galeria de imagens, a navegação na árvore de diretórios, criação de marcadores e exportação dos arquivos.

forensic

Profile de processamento para exames forense genéricos. São habilitados os hashes md5 e sha-256 (necessário para encontrar anexos do WhatsApp), consulta a base de Hashes NSRL do NIST, adicionados e indexados o espaço não alocado e file slacks, e realizado o data carving. Não é feito OCR nem tarefas relativas a casos de pornografia infantil.

pedo

Profile de processamento para exames de casos envolvendo pornografia infantil. São habilitados os hashes md5, sha-1, sha-256 e edonkey, assim são criados marcadores automáticos para itens enviados via Emule, Ares, Shareaza, WhatsApp. São consultadas as bases de hashes NSRL do NIST e de pornografia infantil do LED, bem como é habilitada a detecção de nudez DIE do LED. São indexados o espaço não alocado e fileSlacks. É realizado data carving com uma configuração específica para melhor recuperação de vídeos apagados (o espaço não alocado é dividido em fragmentos de 10GB, sendo que o padrão é 1GB). Também é habilitado o KFFCarving, o qual recupera itens presentes na base de pornografia infantil do LED, mesmo que estejam parcialmente sobrescritos no HD, sem rodapé por exemplo. Além disso é habilitado carving específico para recuperação de itens Known.met do Emule.

blind

Profile para extração automatizada de dados. Essa função deve ser usada com cautela e seu uso indiscriminado não é recomendado. Ela extrai os itens automaticamente com base em categorias pré-definidas no arquivo profiles/blind/conf/CategoriesToExport.txt, tais categorias podem ser adicionadas e removidas pelo usuário.

Durante a extração automatizada, são ignorados itens presentes na bases de hashes NSRL do NIST, é realizado data carving, porém não são indexados o espaço não alocado nem file slacks na configuração padrão. Também o OCR fica desabilitado por padrão, o que pode ser alterado.

Linux

O programa é distribuído com todas as dependências necessárias para execução em ambiente Windows. Para execução em ambiente Linux, é necessário instalar as seguintes dependências:

- Sleuthkit versão 4.4.2 ou superior (baixar do github e compilar atualmente). Sugere-se a aplicação do patch incluído no IPED, permitindo processar os itens antes do término da decodificação da imagem;
- LibEwf, LibVmdk e LibVhdi, para suporte a imagens E01, VMDK e VHD respectivamente;
- Pacote Tesseract versão >= 3.02 e os dicionários português, inglês e OSD (o qual detecta rotação nas imagens)
- Pacote ImageMagick, para permitir a visualização de centenas de formatos de imagens não suportadas pelo Java (o qual decodifica apenas bmp, gif, png e jpg)
- LibreOffice 4 (apenas para análise), para permitir a visualização das dezenas de formatos suportados por essa suite office.
- Oracle Java versão 8 ou superior. É possível utilizar o OpenJDK, porém este não inclui o Webkit do JavaFX, ficando desabilitado o visualizador HTML e derivados (EML, emails de PST, etc)
- MPlayer para geração de miniaturas de vídeos. Recomenda-se a versão 4.9.2. Necessário configurar o caminho para o mplayer no arquivo conf/VideoThumbsConfig.txt.
- Libpff para decodificação de caixas Outlook OST 2013 e melhor decodificação de PSTs, incluindo recuperação de emails apagados. Necessário utilizar versão a partir de 2013 para decodificar OST 2013. Baixar do link Downloads em https://github.com/libyal/libpff/wiki
- Libesedb para expansão de banco de dados EDB (contacts.edb, histórico IE 10 e posteriores, Windows Vista Mail, etc). Baixar de https://github.com/libyal/libesedb/releases
- Módulo perl Parse::Win32Registry para geração de relatórios de registro via RegRipper, já incluído no IPED.

```
perl -MCPAN -e 'install Parse::Win32Registry'
```

- Libmsiecf para decodificação de históricos de internet index.dat do IE 9 e anteriores.

É recomendável desabilitar o swap ou diminuir a tendência do kernel em fazê-lo (swappiness = 10). A maioria das distribuições privilegia o cache de IO e como são lidos mtos gigabytes das imagens, os processos em execução, inclusive o IPED, podem ser paginados para o disco.

No Linux, a execução de processos externos (tesseract, ImageMagick, mplayer, etc) pode copiar parte da memória do processo original durante o fork, o que pode causar problemas quando executada frequentemente a partir de processos que ocupem muita memória (IPED). Por isso, recomenda-se limitar a heap do Java ao executar o processamento: java -Xmx3G ou -Xmx6G são o mínimo e máximo recomendados para um computador com 12 processadores.

DVD bootável com o IPED embutido

O IPED pode ser rodado a partir de um DVD bootável com esta ferramenta embutida baseado na distribuição de Linux forense CAINE, tendo um script na área de trabalho que permite o processamento nos discos rígidos detectados no modo --fastmode. Este DVD já possui os itens destacados acima para rodar de forma completa no Linux, com o Sleuthkit versão 4.3 compilado com os patches do IPED.

Mais informações disponíveis em https://wiki.ditec.pf.gov.br/SEPINF:LED_CAINE.

Utilização

Processamento

Para indexar pastas ou imagens dd, 001, e01 ou iso, utilize a versão linha de comando [1] (https://sepinf.ditec.dpf.gov.br/dav/Softwares/Analise%20de%20Midias/Indexador%20e%20Processador%20de%20Evidências%20Digitais/)

Recomenda-se utilizar um java x64, que permite um maior uso de memória, principalmente em computadores com mais de 4 núcleos de processamento, como as estações periciais de informática PROMOTEC 2.

Uso: java -jar iped.jar -opcao argumento [--opcao_sem_argumento]

Alguns parâmetros são listados abaixo, consulte a ajuda de execução (--help) para verificar todos os parâmetros atualizados.

- -d: dados diversos (pode ser usado varias vezes): pasta, imagem DD, 001, E01, AFF (apenas linux), ISO, disco físico, ou arquivo *.iped (contendo seleção de itens a exportar e reindexar)
- -dname: nome (opcional) para imagens ou discos físicos adicionados via -d
- -o: pasta de saida da indexacao
- -r: pasta do relatorio do AsAP3 ou FTK3
- -l: arquivo com lista de expressoes a serem exibidas na busca. Expressoes sem ocorrencias sao filtradas
- -ocr: aplica OCR apenas na categoria informada. Pode ser usado varias vezes.
- -log: Especifica um arquivo de log diferente do padrao
- -asap: arquivo .asap (Criminalistica) com informacoes para relatorio HTML
- -Xxxx: parâmetros extras de módulos iniciados com -X
- -nocontent: não exporta conteúdo de itens do marcador/categoria informado
- -importkff: importa diretorio com base de hashes no formato NSRL
- -tz: timezone de origem de dispositivos FAT: GMT-3, GMT-4, etc
- -b: tamanho em bytes do setor do dispositivo, necessario informar para discos com setores de 4k

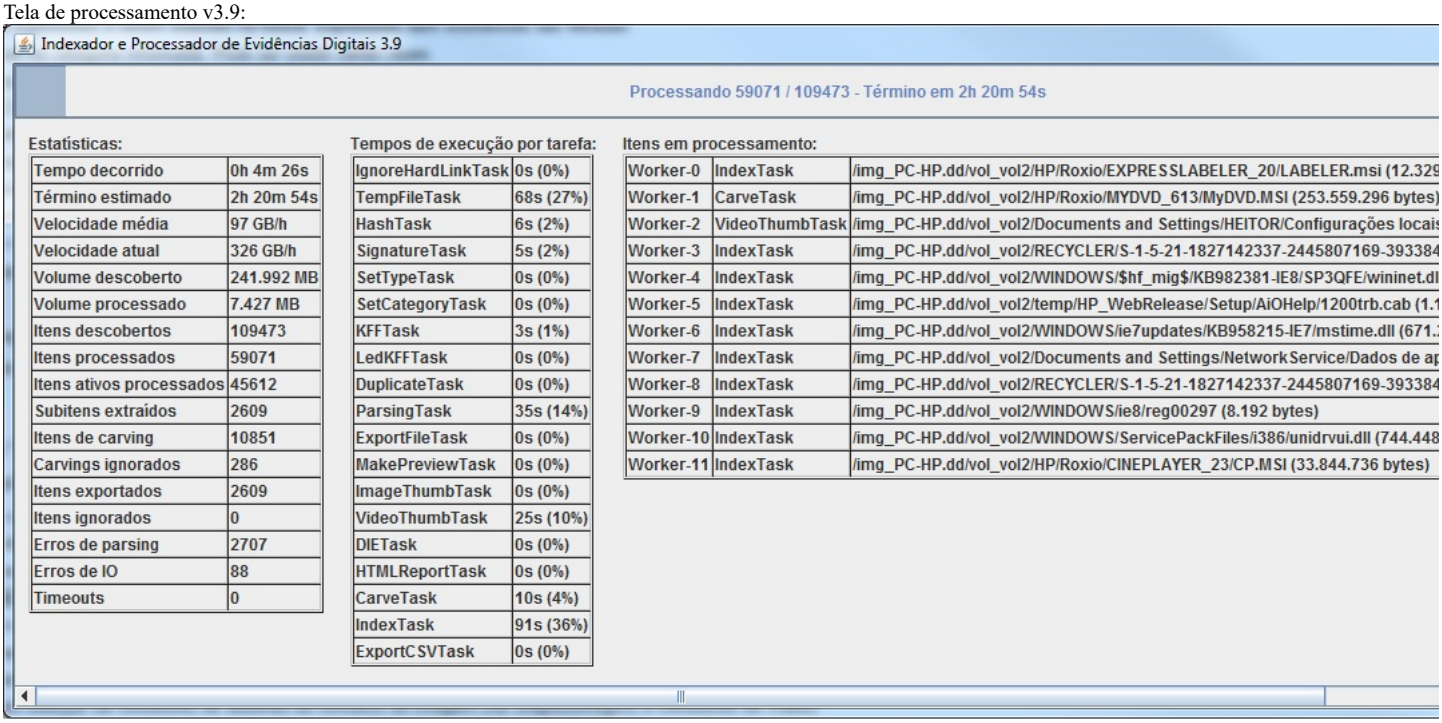
- -profile: usa um profile de processamento: forensic, pedo, fastmode e blind
- --addowner: indexa o owner dos arquivos ao processar pastas (mto lento via rede)
- --append: adiciona indexação a um indice ja existente
- --nogui: nao exibe a janela de progresso da indexacao
- --nologfile: imprime as mensagem de log na saida padrao
- --verbose: gera mensagens de log detalhadas, porem diminui desempenho
- --nopstattachs não exporta automaticamente anexos de emails extraídos de PST/OST

Exemplo:

- java -jar iped.jar -d imagem.dd -o pasta_saída

A versão linha de comando armazena o log em 'IPED/log' enquanto o AsAP 4.4+ o armazena em conf/logs.

Tela de processamento v3.9:



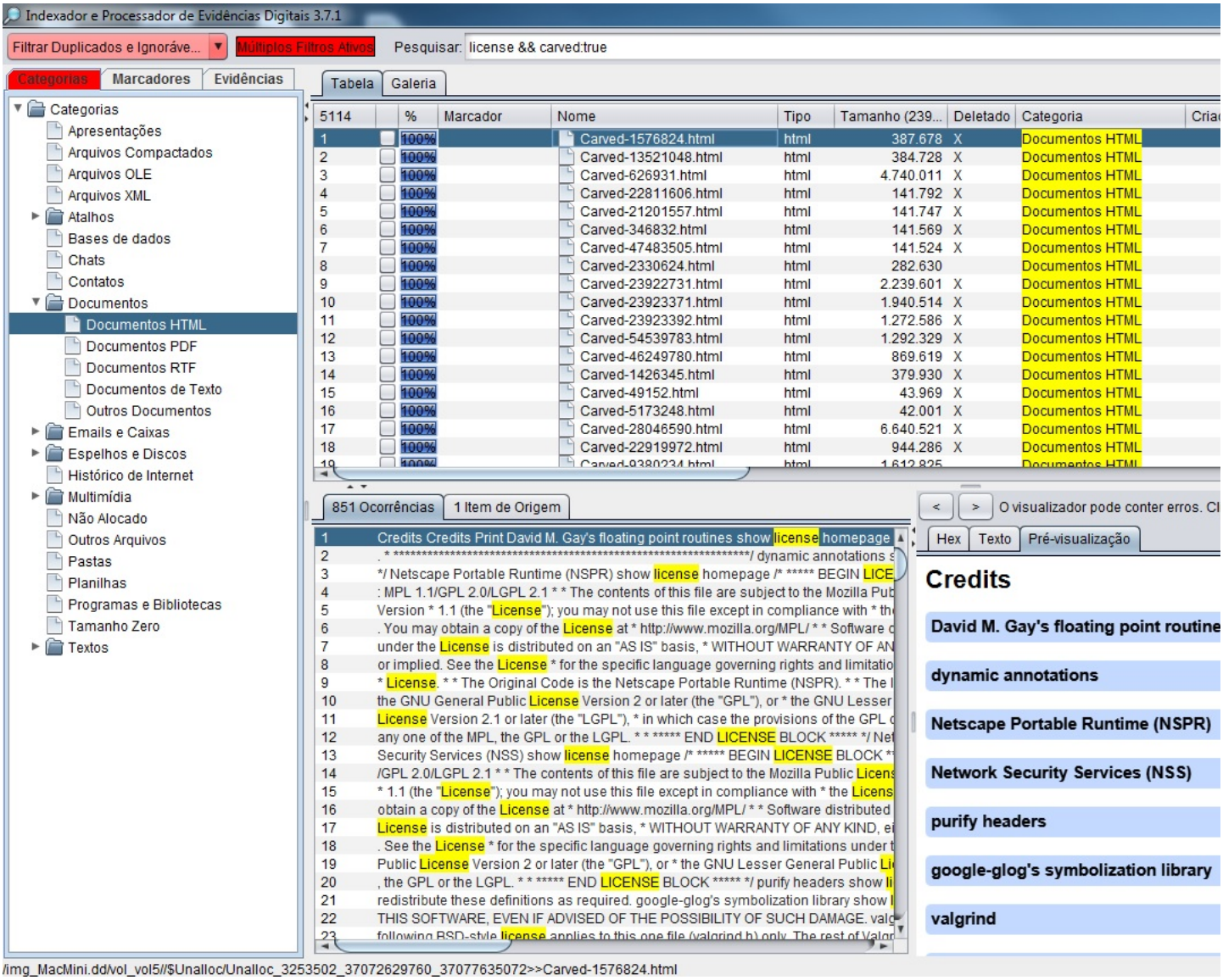
Análise

Acessível pelo executável "Ferramenta de Pesquisa.exe" ou pelo arquivo indexador/lib/iped-search-app.jar. Necessário possuir instalado o Java JRE (recomendado java 64bits), sendo necessário o Java 7 update 06 para habilitar o visualizador Html e EML.

A interface de análise dispõe das seguintes funcionalidades:

- Pesquisa indexada no conteúdo e propriedades dos arquivos
- Painel de fragmentos dos arquivos com ocorrências
- Tabela de resultados ordenável por propriedades
- Visualização em árvore dos dados, recursiva ou não
- Atribuição de múltiplos marcadores textuais, exportação e cópia de propriedades dos arquivos, via menu de contexto
- Galeria multithread para exibição de miniaturas de dezenas de formatos de imagem (via GraphicsMagick) e miniaturas de vídeos
- Visualizador para dezenas de formatos: html, pdf, eml, emlx, rtf, doc, docx, xls, xlsx, ppt, pptx, odt, ods, odp, wps, wpd, sxw, eps, dbf, csv, tif, emf, wmf, odg, pcx, pbm, svg, pict, vsd, psd, cdr, dxf, etc.
- Visualizador de texto filtrado para qualquer formato
- Visualizador hexadecimal
- Exibição dinâmica de colunas, incluindo assunto, remetente e destinatário, autor, cameraModel, data de impressão e centenas de outros.

Versão 3.7.1



Filtros

Por meio da interface de análise é possível realizar filtros de forma intuitiva por palavras-chave, por categoria, por marcador, por diretório, etc. ATENÇÃO, pois sempre são listados na tabela os arquivos resultantes da INTERSEÇÃO de todos os filtros ativos! Caso haja mais de um filtro aplicado, é mostrado um alerta em vermelho para o usuário, para alertar sobre uma possível filtragem de dados acidental.

Além disso, é possível criar e salvar filtros avançados customizados, por meio do botão Opções -> Gerenciar Filtros. Por padrão, são incluídos os seguintes filtros pré-configurados:

- Filtrar Duplicados
- Filtrar KFF Ignoráveis
- Filtrar Duplicados e Ignoráveis
- Alerta de Hash
- Alerta de Hash (PI)
- Arquivos Criptografados
- Erro de Parsing
- Erro de Leitura
- Timeout ao Processar
- Itens Ativos
- Itens Apagados
- Itens Recuperados
- Itens Georreferenciados
- Subitens de Conteineres
- Conteineres não Expandidos
- Imagens com Possível Nudez
- Imagens e Vídeos com Miniaturas

Agrupamento por Metadados

A partir da versão 3.13, foi criada uma nova aba na interface de análise denominada "Metadados". Essa aba é uma generalização da aba de Categorias, pois por meio dela é possível agrupar e filtrar os itens com base nos valores de qualquer metadado ou propriedade existente. Foram criados novos grupos de metadados para facilitar o uso dessa função, por exemplo:

- básicos: nome, extensão, tamanho, categoria, hash, deletado, datas
- avançados: contentType, classeNudez, kffstatus, kffgroup, ocrCharCount, hashes
- email: subject, from, to, cc, bcc, date

- image: model, width, height, date, author, comments
- office: author, product, modified, printed
- regex: cpf, cnpj, email, url, valores, cartao, boleto
- language: detected_1, detected_2, all_detected
- entidades mencionadas: pessoas, empresas, lugares

Podem ser selecionados múltiplos valores do metadado escolhido mantendo Ctrl pressionado. Também é possível ordenar os metadados por número de itens que o possuem, alfabética e numericamente. Para metadados numéricos, é possível escolher uma escala de valores linear ou logarítmica.

IMPORTANTE: Diferente das outras abas, os resultados dessa aba dependem dos filtros aplicados. Isto é, os metadados exibidos se referem apenas aos itens exibidos na tabela (já filtrados) e não a todos os itens do caso. Esse comportamento é proposital para evitar exibir metadados de itens irrelevantes já filtrados. Caso haja alteração nos filtros, basta clicar no botão "atualizar" para atualizar os metadados em exibição.

Georreferenciamento de Imagens

Função incluída na v3.11 pelo PCF Patrick. Basicamente casos contendo imagens com informações de GPS nos metadados exif são renderizadas no painel "Mapa", permitindo visualizar sua localização de origem.

Também foi incluído um filtro "Itens Georreferenciados" para facilitar a localização de itens contendo informações de GPS.

Análise Global de múltiplos casos

Função incluída na v3.11. Com ela é possível abrir e analisar de forma integrada diversos casos simultaneamente, incluindo casos independentes processados em computadores diferentes. Essa função deve ser utilizada via linha de comando no terminal:

```
java -jar lib/iped-search-app.jar -multicases (pasta_com_casos | txt_com_casos)
```

- pasta_com_casos: pasta contendo os casos do iped. Podem aparecer em qualquer nível, não precisando ser filhos imediatos, porém isso pode causar lentidão na inicialização pois é feita uma varredura pelos casos
- txt_com_casos: arquivo txt contendo os caminhos para os casos, um por linha. Os caminhos podem ser absolutos ou relativos a pasta de execução (diretório atual) do usuário

Após a análise global, é possível gerar um relatório único de análise de 2 formas. Uma é salvando um arquivo de marcadores global através do menu "opções -> salvar marcadores", outra é processando os marcadores de cada caso conforme detalhado na seção "Selecionados pelo Perito".

Marcadores Automáticos para Itens Compartilhados

A partir da v3.11, são criados marcadores automáticos para itens localizados na evidência e compartilhados via aplicativos Emule, Ares e Shareaza, e a partir da v3.12, para itens enviados via Skype e WhatsApp. Para o funcionamento é necessário, antes do processamento, habilitar os hashes utilizados por cada aplicativo no controle de arquivos transferidos: md5 para Shareaza, sha-1 para Ares, edonkey para Emule, sha-256 para WhatsApp. Basicamente os hashes presentes nos arquivos de controle de transferências são pesquisados no caso e, caso encontrados, são criados marcadores automáticos para esses itens por aplicativo. Tais itens não são necessariamente ilícitos, devendo ser inspecionados pelo perito. No caso do Skype, não há registro dos hashes dos itens transferidos, por isso são utilizadas as informações de "tamanho" E "caminho original", ou "tamanho" E "nome" caso o item não seja localizado no "caminho original" ou caso esta informação não esteja disponível.

Os marcadores criados são automáticos e referem-se a transferências prováveis, devendo ser confirmadas pelo perito.

Localização de documentos por similaridade

Na v3.12, foi incluído no menu Opções a função "Encontrar documentos similares". Com ela é possível localizar documentos textuais que tenham conteúdo ou assunto similar ao documento em foco, de acordo com um porcentual de similaridade informado pelo usuário.

É utilizada uma heurística para identificar palavras representativas dos documentos e, de forma simplificada, quanto mais palavras representativas em comum, mais similares são considerados os documentos.

Análise Simultânea ao Processamento

Função incluída na versão 3.12, possibilitando analisar o caso antes do término do processamento. A qualquer momento pode ser acionado o botão "Atualizar" na interface, disponível apenas durante o processamento, para carregar novos itens processados.

Importante destacar que, atualmente, apenas ficam disponíveis para análise itens processados por completo. Itens com alguma tarefa pendente (como indexação) ficam indisponíveis, inclusive na árvore de diretórios, a qual fica incompleta até o término do processamento. Porém, itens já disponibilizados para análise estão finalizados, significando que todas as funções de análise já estão disponíveis para esses itens, como visualização de miniaturas de imagens e vídeos na galeria, filtro de itens sem miniaturas, buscas indexadas, georreferenciamento, etc. Além disso, os itens não mudarão de categoria como em outros softwares, pois a análise de assinatura, assim como as demais tarefas, já terá sido finalizada.

Extração de Arquivos de Interesse

Automática

Essa funcionalidade deve ser utilizada com cautela e seu uso indiscriminado não é recomendado.

- Por Categorias:

Para realizar extração automática de arquivos por categoria, utilize a opção "-profile blind" ou descomente as categorias de interesse no arquivo conf/CategoriesToExport.txt antes do processamento. Nesse caso, os arquivos são exportados e apenas eles indexados, podendo o resultado do processamento ser enviado para o solicitante da extração de dados, via mídia óptica ou magnética.

- Por Palavras-chave:

Para realizar extração automática de dados por palavras-chave, insira as palavras-chave ou expressões regulares de interesse, uma por linha, no arquivo conf/KeywordsToExport.txt dentro do profile desejado. Os arquivos contendo as palavras ou expressões definidas serão exportados. Diferenças de capitalização e acentuação são ignoradas. Os hits são adicionadas ao metadado Regex:KEYWORDS dos itens, permitindo filtrar os itens por palavra/expressão encontrada.

Caso sejam definidas categorias e palavras-chave a exportar automaticamente, é feito um OR do resultado de cada exportação configurada.

Selecionados pelo Perito

Após atribuir marcadores aos arquivos (opcional), os arquivos de interesse a serem exportados devem ser selecionados (via checkbox) na interface de análise. Essa seleção pode ser salva via função "Salvar marcadores" em um arquivo *.iped, ex: Report.iped. Posteriormente, via terminal, forneça esse arquivo (ou o arquivo padrão indexador/marcadores.iped) como parâmetro para uma nova indexação:

```
java -jar iped.jar -d Report.iped -o pasta_relatorio
[[-nocontent "Marcador1"] [-nocontent "Marcador2] ...]
```

- -nocontent Marcador1: exporta as propriedades e as miniaturas dos itens do Marcador1, mas nao exporta os arquivos originais, útil para imagens e vídeos.

A partir da versão 3.11, é possível gerar um único relatório a partir de múltiplos casos:

```
java -jar iped.jar -d marcadores1.iped -d marcadores2.iped -o pasta_relatorio_unico
```

Também pode ser usada a opção --nopstattachs caso se queira desabilitar a inclusão automática no relatório de anexos de emails selecionados extraídos de PST.

Os arquivos selecionados serão exportados e reindexados para facilitar a revisão dos dados pelo solicitante do exame.

Relatório HTML

(contribuição PCF Wladimir)

A partir da versão 3.4, no caso de extração de arquivos de interesse (automática ou de selecionados), por padrão é gerado um relatório HTML com propriedades e links para os arquivos extraídos. Nesse relatório é incluída uma galeria de imagens e também miniaturas dos vídeos (caso geradas). Também são geradas versões de visualização para itens com parser apropriado. Alguns parâmetros do relatório podem ser alterados no arquivo conf/HTMLReportConfig.txt

Consumo de Memória e Performance

No caso de problemas de falta de memória, aumente a memória heap do java (-Xms), diminua o parâmetro "numthreads" no arquivo IPEDConfig.txt ou, preferencialmente, utilize um Java 64bits. Recomenda-se utilizar uma versão x64 do java, que permite um maior uso de memória, quando necessário, pois a heap padrão do java x86 é de apenas 256MB, insuficiente para processar imagens.

Por padrão (numThreads = default), cada processador lógico executa uma thread de processamento, que geralmente consome 250MB de memória (max de ~500 MB). É recomendado limitar a memória utilizada pelo java, pois a partir de certo ponto adicionar mais memória à aplicação não adianta, muito melhor é deixar memória livre para ser utilizada como cache de IO. Pode-se utilizar o parâmetro -Xmx para limitar a memória heap da JVM. Uma regra simples é configurar -Xmx com metade da memória RAM disponível, ou utilizar a regra citada: num_processadores*512MB. Por exemplo, numa máquina com 12 núcleos: java -Xmx6G -jar iped.jar -d imagem.dd -o pasta_saida.

O processamento completo (assinatura, hash, expansão, indexação e carving) geralmente fica em torno de 100 GB/h a 300 GB/h na estação HP820. Quando não habilitado o OCR, geralmente o gargalo é o I/O do disco que contém a imagem sendo processada.

Recomenda-se configurar uma pasta de saída (-o) do processamento num disco diferente daquele que contém os dados sendo processados, para minimizar acessos concorrentes de leitura dos dados e escrita de subitens expandidos.

Sempre que possível configure o diretório temporário (indexTemp no arquivo IPEDConfig.txt) num disco rápido, diferente daquele que contém os dados, fora do disco de sistema e livre de antivírus, preferencialmente num SSD. Também indique se o indexTemp encontra-se num disco SSD ou não (indexTempOnSSD). Caso indicado, são feitas otimizações que podem diminuir o tempo de processamento para menos da metade: o número de threads de merges do índice é aumentado e, no caso de imagens compactadas E01, são gerados arquivos temporários para todos os itens para evitar múltiplas descompactações dos itens pela LIBEWF, a qual não é multithread e efetua apenas uma descompactação por vez, subaproveitando processadores com vários núcleos.

Também é altamente recomendado configurar a base KFF num disco SSD, sob pena de impactar severamente o tempo de processamento.

O programa foi otimizado para suportar casos na casa de 10 milhões de itens com poucos giga de memória, havendo degradação principalmente na ordenação das propriedades atualmente.

Considerações finais

A precisão dos resultados têm sido considerada satisfatória para atender aos objetivos propostos, mas nunca será 100%, pois há uma infinidade de tipos de arquivos a tratar. Por isso, os resultados da ferramenta podem diferir dos resultados de outras ferramentas forenses, podendo haver diferença tanto para mais quanto para menos. Por exemplo, em relação a ferramenta AccessData FTK, atualmente há diferenças de configuração que resultam num menor número de itens no caso, o que pode ser equivocadamente interpretado como uma deficiência do software numa análise superficial. Note que a inclusão de muitos itens inúteis no caso pode dificultar a análise ao invés de ajudar. Abaixo são citadas algumas das diferenças:

- trechos não alocados são menos fragmentados;
- arquivos de carving claramente corrompidos, menores que 1KB ou maiores que 10MB são ignorados na configuração padrão;
- não são expandidos históricos de Internet nem arquivos JAR (que podem produzir dezenas de milhares de itens .class);
- informações EXIF são extraídas sem a criação de subitens sob as imagens;
- não são gerados itens desnecessários com o resultado do OCR
- não são extraídas miniaturas das imagens a partir da v3.6

Caso sejam identificadas divergências importantes, entre em contato e relate o problema para que ele possa ser corrigido.

No caso de dúvidas, bugs ou sugestões, envie um e-mail para nassif.lfcn@dpf.gov.br

---