Demo: Creating a Block Hash Database

Using *hashdb* and *md5deep*

We build large databases of block hashes to help us find fragments of previously encountered data. Here are some ways *hashdb* databases are used:

- We scan data looking for block hashes that match block hashes in previously encountered data.
- We subtract data known to be not interesting, such as system files, to remove distracting false positives.

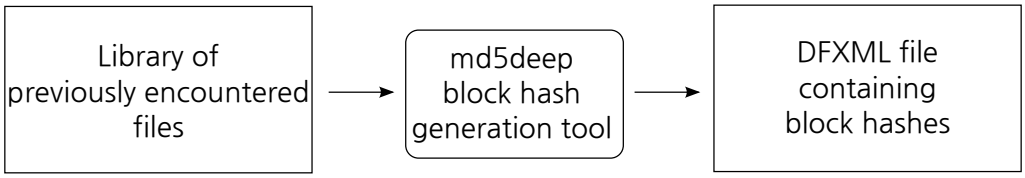
The Scanner demo at `http://digitalcorpora.org/downloads/hashdb/demo/scanner_demo.pdf` uses a block hash database to find previously encountered data in a media image.

The Similarity demo at `http://digitalcorpora.org/downloads/hashdb/demo/similarity_demo.pdf` compares two block hash databases created from two media images to find fragments of user data that are common between them.

In this demo, we create block hash database `mock_video.hdb` of mock video file `mock_video.mp4` to use as a reference for finding previously encountered data. The database we create is the very database we use in the introductory demo for finding fragments of previously encontered data at `http://digitalcorpora.org/downloads/hashdb/demo/scanner_demo.pdf`.

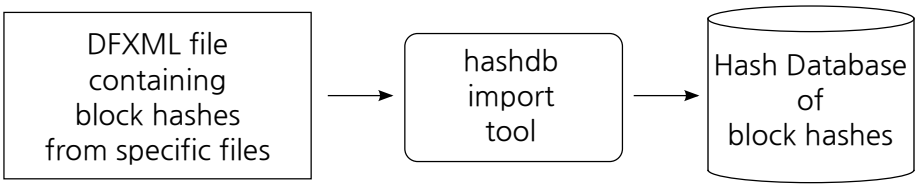
Workflow:

1. Generate a DFXML file containing block hashes from sources of "previously encountered" data.



Run *md5deep* to create a DFXML file of block hashes from your library of prevously encountered files.

2. Import the DFXML file of block hashes into a block hash database.



Run the *hashdb* `import` command to import block hashes into the hash database.

Steps:

1. Download and install *hashdb* from `http://digitalcorpora.org/downloads/hashdb` as described at `https://github.com/simsong/hashdb/wiki/Installing-hashdb`.
2. Download and install *bulk_extractor* compiled with *hashdb* from `http://digitalcorpora.org/downloads/hashdb` as described at `https://github.com/simsong/hashdb/wiki/Installing-hashdb`.
3. Download and install the *md5deep* tool as described at `https://github.com/simsong/hashdb/wiki/Installing-hashdb`.
4. Pick a directory or file to use as your source of files. For this demo, please download and use the mock video file at `http://digitalcorpora.org/downloads/hashdb/demo/mock_video.mp4`.
5. Use the *md5deep* tool to generate the DFXML file of block hashes of your files and directories:
   - Use `-p 4096` to specify a block (partition) size of 4096.
   - Use `-d` to generate output in DFXML format.
   - Use `> mock_video_hashes.xml` to direct the DFXML output to go to file `mock_video_hashes.xml`.

   ```
   $ md5deep -p 4096 -d mock_video.mp4 > mock_video_hashes.xml
   ```
6. Use the *hashdb* tool to `create` a new hash database called `mock_video.hdb`:
   ```
   $ hashdb create mock_video.hdb
   ```
7. Use the *hashdb* tool to `import` the block hashes from the DFXML file into the new hash database:
   ```
   $ hashdb import mock_video.hdb mock_video_hashes.xml
   ```

This completes the demo.