


JOHANNES KEPLER
UNIVERSITY LINZ | JKU

# Automating Web History Analysis

**Michael Sonntag**

Institute for Information processing and microprocessor technology (FIM)

Johannes Kepler University Linz, Austria

sonntag@fim.uni-linz.ac.at

# Why?

- Separate applications in the Internet are dying out
- Almost everything takes place (or is being moved) to the WWW
  - Exception: Apps on mobile devices
- Examples:
  - Social networks: "Native" web applications
  - Mail clients & calendars: Client program → Web-based interface
  - File storage: FTP → Sharehoster
  - Office software (documents, spreadsheets): Programs → Online services
  - Listening to music: Downloading → Streaming services (only interface via WWW!)
- Result: Web history gets more important, but also increasingly large & complex
  - Not a single browser window only, but multiple tabs + other applications

# Why “we”?

- Especially communication and collaboration has changed
  - Mandatory data retention: Don’t use E-Mail → Web services!
  - Typical mode of communication for terrorists:
    - Draft E-Mail in web-based E-Mail system and store it, second person logs in and reads this same draft (and perhaps modifies it or generates another draft)
  - Typical exchange of illegal data:
    - Encrypt, assign random filename, upload to sharehoster
    - Pass on link in some way (perhaps: web forums)
    - Pass on password in some way (chat systems)
- All of these leave very few traces, and transport is encrypted (TLS!)
  - Inspect the local web history to find traces of this behaviour

# What?

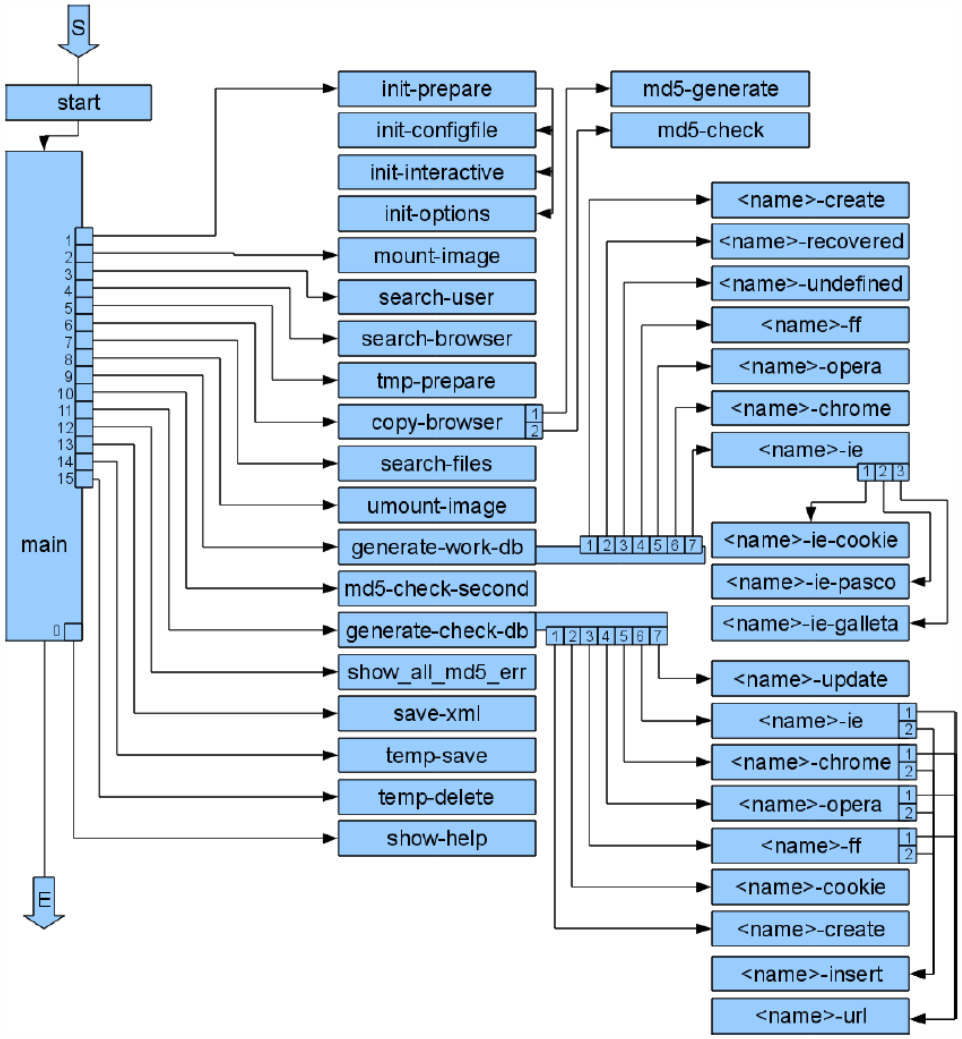
- Identifying the behaviour of users, based on their web activity traces
  - Traces remaining on the computer itself, i.e. no computer → no evidence!
    - Not: Data retention, online search!
- As much details as possible:
  - Which web pages were visited and what was their content?
  - Intentionality: Clicking or automatic JavaScript/PopUp?
  - Hints to external servers: Webmail, file storage etc.
  - Additional information: User names/passwords, visit count, …
- From all browsers: A user might employ several, even on the same system
- Result: SW for automatically collecting and collating all this information
  - History, cache, and cookies → Timeline + sessions

# How?

- Command line tool: Important for automating the process
- Results: Text, CSV, XML, SQLite-DB → For extraction and evaluation
- Basic input: (Forensic – hopefully!) disk image
  - Not a copy of the files, but of all sectors of the disk → Requires mounting of file system
- Restriction to selected users possible if desired (useful for large servers)
- Reconstruction of deleted files optionally possible through external tool
- Validation of cache (+help for collation) against current web content (takes long!)
- Hash values are created & verified to protect against modifications
- Special handling of each type of browser → Simple extensibility
  - Supported: IE (requires two open source tools), Firefox, Chrome, Opera

Implementation: Gerald Prock (Master thesis)

# How?


S
start
main
E
init-prepare
init-configfile
init-interactive
init-options
mount-image
search-user
search-browser
tmp-prepare
copy-browser
search-files
umount-image
generate-work-db
md5-check-second
generate-check-db
show_all_md5_err
save-xml
temp-save
temp-delete
show-help
md5-generate
md5-check
<name>-create
<name>-recovered
<name>-undefined
<name>-ff
<name>-opera
<name>-chrome
<name>-ie
<name>-ie-cookie
<name>-ie-pasco
<name>-ie-galleta
<name>-update
<name>-ie
<name>-chrome
<name>-opera
<name>-ff
<name>-cookie
<name>-create
<name>-insert
<name>-url

# Show me!

- Exemplary output (XML version) of a single HTML file in the cache:

```
<entry>
    <id>43</id>                                                    Internal number
    <systemname>windows_new</systemname>                           From which computer the data was collected
    <username>MSCWin7</username>                                   Username of the inspected user
    <browsername>ie</browsername>                                  Which browser
    <type>http</type>                                              http (=cache)/https (=cache)/cookie/visited(=history)/…
    <url>http://p8.friendscout24.de/e5e892b8a6eeb710c45b168bc5f2bd2a/0/78/57/29/14/74C3F09B5B4194909AD3FDBE820
    2E142s.jpg</url>                                               The URL of this entry
    <title>://p8.friendscout24.de/e5e892b8a6eeb710c45b168bc5f2bd2a/0/78/57/29/14/74C3F09B5B4194909AD3FDBE8202E
    142s.jpg           09/12/2009 19:24:09     09/12/2010 19:26:20
           74C3F09B5B4194909AD3FDBE8202E142s[1].jpg        MSKM6NQ5    HTTP/1.0 200 OK  Pragma: cached
    Content-Type: image/jpeg  Content-Length: 2123  X-Cache: HIT from p.friendscout24.de    ~U:mscwin7</title>
                                                                   The stored HTTP response header
    <from_visit>none</from_visit>                                  ID of the previous URL
    <last_visit_time>1284312380</last_visit_time>                  The time of the last visit of this URL
    <visit_count>0</visit_count>                                   How often visited (0 = Not available)
    <typed_count>0</typed_count>                                   How often manually typed (0 = Clicked on link/Bookmark/…)
    <hidden>0</hidden>                                             Loaded in background (0 = not preloaded)
    <test_session_number>1</test_session_number>                   Session number of reachability testing
    <test_session_time_start>1284312381</test_session_time_start>        Start/End of testing session
    <test_session_time_stop>1284312783</test_session_time_stop>
    <test_cache_exist>1</test_cache_exist>                         File has been found in the cache (here obvious!)
    <test_cache_same>0</test_cache_same>                           File in cache differs from the one loaded in testing
    <test_url_exist>1</test_url_exist>                             URL was found in testing (some page exists there)
    <test_url_title_same>0</test_url_title_same>                   Title of test page download differs
    <test_cookies_set>0</test_cookies_set>                         Test session did not set any cookies
    <test_cookies_exist>0</test_cookies_exist>                     No cookies from test session found in cookies folder
    <test_cookies_time_ok>0</test_cookies_time_ok>                 Number of cookies where the timestamp is similar to the
                                                                         ones downloaded during testing
</entry>
```

# What next?

- Adding further browsers, e.g. Safari, and especially those of mobile devices
  - Typically Webkit-based, so similar to Chrome/Firefox and each other
- Extracting more information: Difficult, as available data varies significantly between the browsers
- Automating the integration of several runs of the software
  - Several devices/computers of a single person
- Support for interpretation: Visualization, statistics, comparison to lists of known good/bad data (e.g. URLs or page content), search functionality (full-text index)
- Preprocessing: Assigning cookies to pages (note: need not be same domain → cache parsing!), web archive integration to complete missing cache elements, …

# Conclusions

- Helpful program getting rid of very tedious and uninteresting work
  - Especially dealing with images and hash values: Necessary for forensic work!
- But still a lot of things to do: Interpretation
  - What does this mean? What did the person do? Anything in there suspicious?
    - Looking whether a specific URL was visited is trivial, but often the behavior is not nearly as easy to detect or conclusive!
  - Integration with other tools is still an effort, but at least it becomes possible
    - Existing tools often have GUI-output only!
- Sufficient for interpreting "web browsing"; doesn't work for web services
  - No "page" concept exists, not stored in "browser data"
    - Requires inspecting the web traffic itself (wiretapping)

# Thank you for your attention!

**Michael Sonntag**

Institute for Information processing and microprocessor technology (FIM)

Johannes Kepler University Linz, Austria

sonntag@fim.uni-linz.ac.at