

Forensic Analysis of Web Data

Erhard Dinhobl

Vienna University of Technology - Information and Software Engineering Group

3.10.2016

Introduction

- ▶ using collected data in surveys
- ▶ influence to known investigative models
- ▶ appropriate investigation model
- ▶ digital preservation
- ▶ legal issues
- ▶ deep web crawling
- ▶ format of data
- ▶ designing a deep web crawler

Table of contents

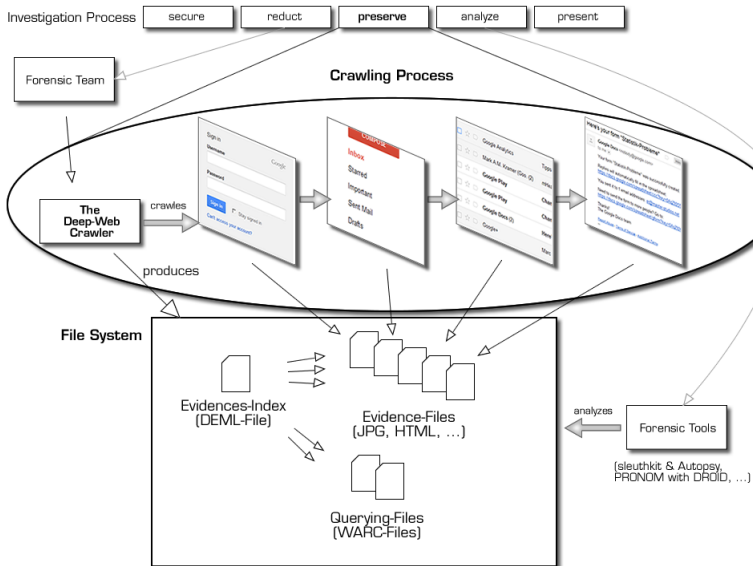
- ▶ base investigation model
- ▶ new investigation model
- ▶ "reduct", "preserve" and "analyze" steps in detail
- ▶ underlying methodology (deep web crawler)
- ▶ usage

The Investigation S(RP)AP

- ▶ secure
- ▶ **reduct**
- ▶ **preserve**
- ▶ analyze
- ▶ present

Best for online crime investigation!

The Process



The Process (2)

- ▶ Where was the evidence stored?
- ▶ Who had obtained the evidence?
- ▶ What has been done to the evidence?

"reduct"

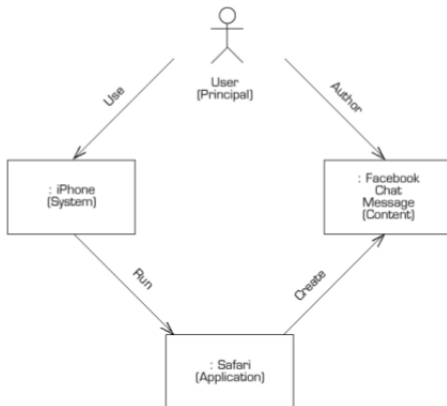
- ▶ reduce amount of data to crawl (also time)
- ▶ crawling websites: XPath
- ▶ simulate user navigation and interaction
- ▶ defined by investigation team
- ▶ be aware: unimportant data maybe later important

"preserve"

- ▶ preservation of evidences up to 60 years (StPO §60)
- ▶ important due to file formats (determined with DROID)
- ▶ for web crawling WARC file format (open format)
- ▶ for immediate investigation: normal files
- ▶ the index: DEML file
 - ▶ by National Centre for Forensic Science
 - ▶ Digital Evidence Markup Language
 - ▶ compatible with Global Justice XML Data Model (GJXDM)
 - ▶ File (inheritance: DigitalArtefact, Allocated) important for online investigations

"analyze"

- ▶ file analysis: sleuthkit, Autopsy, PRONOM (with DROID)
- ▶ own tools and usage of graph-expression (Google)
- ▶ timeline and correlation reconstruction:
 - ▶ model from the National Centre for Forensic Science
 - ▶ Application, Content, Principal, System

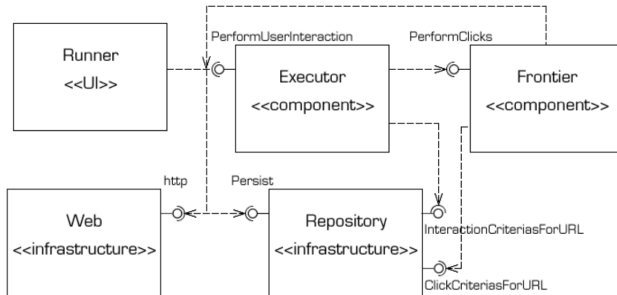


The Deep Web Crawler

Challenges:

- ▶ the use of AJAX technology in web pages
- ▶ the simulation of user interaction with HTML elements (e.g. forms)
- ▶ the use of Captchas in web pages

The Architecture



Simple Example

```
<?xml version="1.0"?>
<cratalis start-url="http://kurier.at"
  name="Kurier Crawling"
  start-action-sequence="1">
  <action-sequence id="1">
    <action type="StartAction" id="1"
      comment="start every link" save="true"
      action-sequence-id="3">
      .//a[@href and
        not(starts-with(@href, "#"))]
    </action>
  </action-sequence>
  <action-sequence id="3">
    <action type="SaveAction" id="3"
      comment="just save" save="true"/>
  </action-sequence>
</cratalis>
```

More complex example

```
<?xml version="1.0"?>
<cratalis start-url="http://www.studivz.de" name="StudiVZ" start-action-sequence="1"
  deml-file="studivz/studivz.xml">
  <action-sequence id="1">
    <action type="JavaScript" id="1" comment="filling out the login form"
      save="true" save-file="studivz/login-screen.html">
      var usernameInputs = document.evaluate(".*[@id='Login_email ']", ...);
      var usernameTextBox = usernameInputs.iterateNext();
      usernameTextBox.value = "forensicanalysis@ymail.com";
      var passwordInputs = document.evaluate(".*[@id='Login-password ']", ...);
      var passwordTextBox = passwordInputs.iterateNext();
      passwordTextBox.value = "gustav";
    </action>
    <action type="StartAction" id="2" save-file="studivz/home.html"....></action>
  </action-sequence>
  <action-sequence id="2">
    <action type="StartAction" id="3" save-file="studivz/inbox.html"></action>
  </action-sequence>
  <action-sequence id="3">
    <action type="JavaScript" id="4" comment="start other action to overview" ...>
      var msgLinkList = document.evaluate(
        ".*[div[contains(@class,'from-subject')]/.../a", ...);
      var msgLink;
      var count = 0;
      while(msgLink = msgLinkList.iterateNext()) {
        count++;
        msgLink.onclick.apply(msgLink);
        savePage('studivz/msg' + count + '.html');
      }
    </action>
  </action-sequence>
</cratalis>
```

Usage

- ▶ investigative
 - ▶ collecting bitcoin addresses (clear and dark net)
 - ▶ saving evidence files from the web (also via Tor)
- ▶ non-investigative
 - ▶ collecting event data
 - ▶ ... market watch
 - ▶ ... person data (incl. relations)
 - ▶ ... news
 - ▶ ... locations
 - ▶ ... bitcoin addresses and names
 - ▶ many others still done