# Outils de traitements de logs Apache

1) Anonymisation des logs

2) Outil visuel d'exploration des données

3) Adaptation d'un robot



*Guyet Thomas, INRIA/IRISA, Equipe DREAM, le 05/05/08*

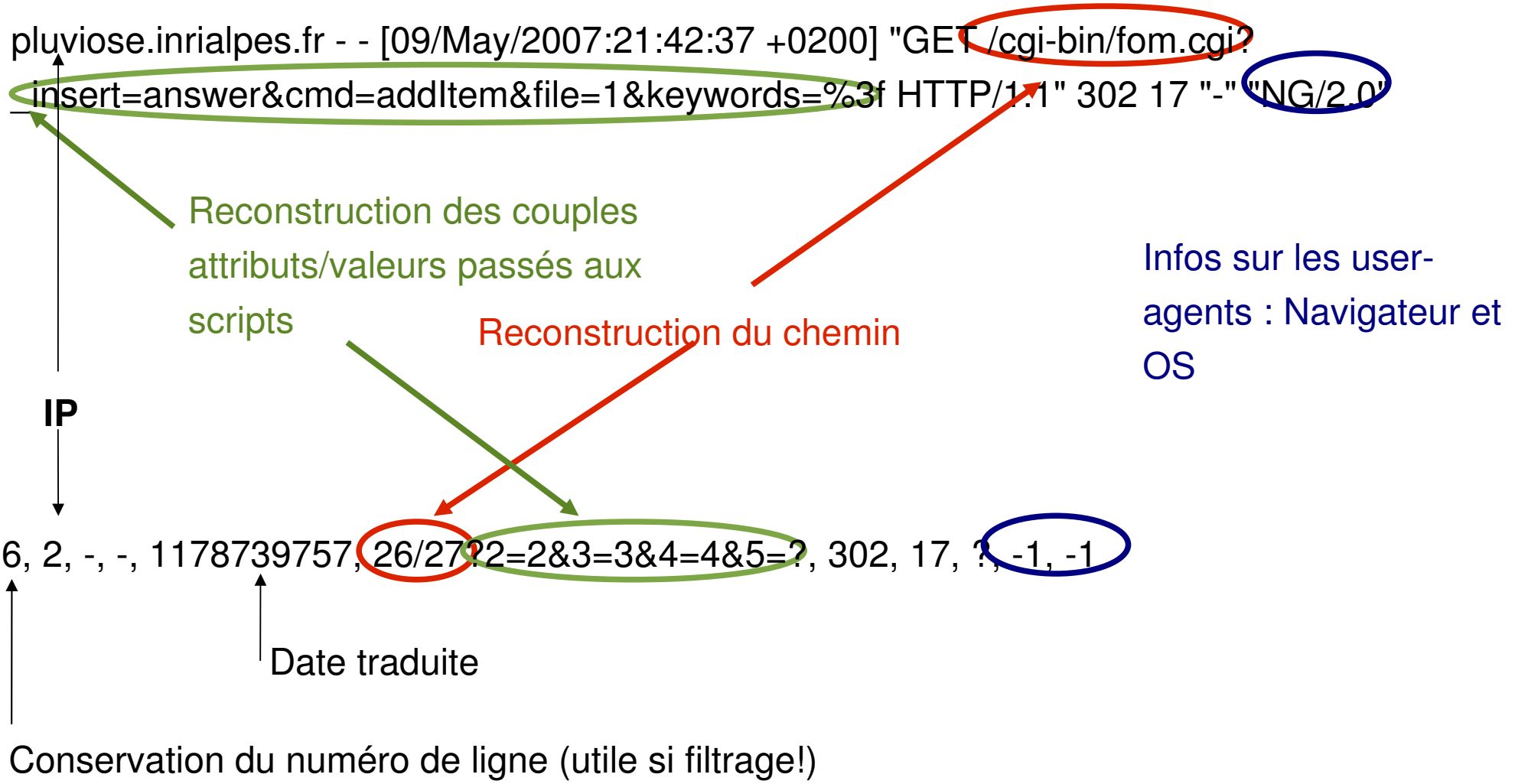
# Anonymisation des logs

- Objectifs :
  - Anonymiser les logs
    - du point de vue des clients (utilisateurs) : IP
    - du point de vue du serveur (données sur le serveurs) : les requêtes
  - Conserver une information permettant l'exploitation des logs
    - Difficile de savoir *a priori* ce qu'il faut conserver
    - Donc, on cherche à en conserver le maximum

- Outil fonctionnant en ligne de commandes

```
./anonimizer -o anonymizedlogfile logfile
```

# Anonymisation des logs



# Anonymisation des logs

crawl-66-249-66-136.googlebot.com - - [09/May/2007:21:42:39 +0200] "GET /sloop/David.Coudert/Biblio/Year/index.php3?url=Biblio/Year/2005.complete.html HTTP/1.1" 200 176 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
pluviose.inrialpes.fr - - [09/May/2007:21:42:27 +0200] "GET /axis/cbrtools/manual/first_page.html HTTP/1.1" 304 - "-" "NG/2.0"
eruessel.stusta.mhn.de - - [09/May/2007:21:42:27 +0200] "GET /lemme/Hanane.Naciri/these/mathml/images/box.gif HTTP/1.0" 200 183 "http://www-sop.inria.fr/lemme/Hanane.Naciri/these/mathml/main.html" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3"
eruessel.stusta.mhn.de - - [09/May/2007:21:42:27 +0200] "GET /lemme/Hanane.Naciri/these/mathml/images/archmathml.gif HTTP/1.0" 200 6665 "http://www-sop.inria.fr/lemme/Hanane.Naciri/these/mathml/main.html" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3"
crawl-66-249-66-136.googlebot.com - - [09/May/2007:21:42:29 +0200] "GET /reves/Xavier.Granier/GIS/html/class_clusterizer.html HTTP/1.1" 200 7602 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
pluviose.inrialpes.fr - - [09/May/2007:21:42:37 +0200] "GET /cgi-bin/fom.cgi?_insert=answer&cmd=addItem&file=1&keywords=%3f HTTP/1.1" 302 17 "-" "NG/2.0"
pluviose.inrialpes.fr - - [09/May/2007:21:42:37 +0200] "GET /cgi-bin/fom.cgi?_insert=answer&cmd=addItem&file=1&keywords=%3f HTTP/1.1" 302 17 "-" "NG/2.0"
ferrier.biac.duke.edu - - [09/May/2007:21:43:56 +0200] "GET /asclepios/personnel/Pierre.Fillard/softwares/FiberTracking/DTITrack2005_manual.pdf HTTP/1.1" 206 585350 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 2.0.50727; InfoPath.1)"
crawl-66-249-66-136.googlebot.com - - [09/May/2007:21:42:27 +0200] "GET /acacia/ESSI/Images/%3FS=A&h=437&w=822&sz=9&hl=en&start=7/Fig-cycle-conception-IHM-LN.fm/Bad-printer-icon.fm/proprietes-IHM-LN.ps/osf.TIFF.gz/Logo-Moduel-IHM-design.ppt/arbre-des-colecticiels/Bad-printer-icon.fm/Bad-printer-icon.fm/Bad-printer-icon.ps/Fig-Modele-activite-Norman.fm/?D=A HTTP/1.1" 200 11880 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)"
eruessel.stusta.mhn.de - - [09/May/2007:21:42:27 +0200] "GET /lemme/Hanane.Naciri/these/mathml/images/figue.gif HTTP/1.0" 200 1540 "http://www-sop.inria.fr/lemme/Hanane.Naciri/these/mathml/main.html" "Mozilla/5.0 (Windows; U; Windows NT 5.1; de; rv:1.8.1.3) Gecko/20070309 Firefox/2.0.0.3"

# Anonymisation des logs

1, 1, -, -, 1178739759, 1/2/3/4/5?1=1, 200, 176, ?, -1, 1
2, 2, -, -, 1178739747, 6/7/8/9, 304, 0, ?, -1, -1
3, 3, -, -, 1178739747, 10/11/12/13/8/5, 200, 183, 14/15/16/17/18/19/20, 1, 3
4, 3, -, -, 1178739747, 10/11/12/13/8/21, 200, 6665, 14/15/16/17/18/19/20, 1, 3
5, 1, -, -, 1178739749, 22/22/23/24/25, 200, 7602, ?, -1, 1
6, 2, -, -, 1178739757, 26/27?2=2&3=3&4=4&5=?, 302, 17, ?, -1, -1
7, 2, -, -, 1178739757, 26/27?2=2&3=3&4=4&5=?, 302, 17, ?, -1, -1
8, 4, -, -, 1178739836, 28/29/8/13/30/31, 206, 585350, ?, 1, 2
9, 1, -, -, 1178739747, 32/33/34/?6=5&7=6&8=7&9=8&10=9&11=?12=5, 200, 11880, ?, -1, 1
10, 3, -, -, 1178739747, 10/11/12/13/8/35, 200, 1540, 14/15/16/17/18/19/20, 1, 3

# Anonymisation des logs

- Les tables (4 fichiers)

**IPs**

1: crawl-66-249-66-136.googlebot.com
2: pluviose.inrialpes.fr
3: eruessel.stusta.mhn.de
4: ferrier.biac.duke.edu

**Valeurs**

1: BiblioYear2005.complete.html
2: answer
3: addItem
4: 1
5: A
6: 437
7: 822
8: 9
9: en

**Attributs**

1: url
2: _insert
3: cmd
4: file
5: keywords
6: S
7: h
8: w
9: sz
10: hl
11: start
12: D

**Chemins**

1: sloop
2: David.Coudert
3: Biblio
4: Year
5: index.php3
6: axis
7: cbrtools
8: manual
9: first_page.html
10: lemme
11: Hanane.Naciri
12: these
13: mathml
14: ...

# Les options (1)

- Choix du format de sortie
  - Éléments de la ligne de log à conserver
  - Permet la suppression d'éléments non-utilisés : *hyphen*, *userid*, ...

- Options sur les formats d'anonymisation
  - Date traduite en entier ou non
  - 3 formats d'anonymisation des requêtes :
    - chemin et attributs/valeurs,
    - sans tenir compte des attributs/valeurs,
    - globalement : 1 identifiant pour chaque requête

# Les options (2)

- Réutilisation de tables existantes
  - Permet de traduire un nouveau fichier de logs en utilisant les mêmes identifiants

- Filtrage *a priori* des logs

# Filtrage des logs

- Objectif : Ne proposer que les lignes de logs « interessantes » à explorer
  - Filtrer les robots
  - Filtrer les requêtes à des pages statiques pour l'étude des intrusions

- Utilisation de filtres
  - Un filtre = une expression régulière
  - Appliqué sur IPs, requêtes ou user-agents

# Filtrage des logs

- Possibilité de charger un fichier de filtres

```
# Some Googlebot filter based on IP
IP 1 crawl-66-249-66-[0-9]{1,3}\.googlebot\.com
IP 1 livebot
IP 1 natcrawlbloc

# Crawler filter based on the UserAgent (UA) string
UA 1 Slurp
UA 1 msnbot

# Filter static pages
RE 1 html$
RE 1 .html#
RE 1 htm$
RE 1 xml$
```

# Filtrage de logs

6, 1, -, -, 1178739757, 11/12?1=1&2=2&3=3&4=?, 302, 17, ?, -1, -1

7, 1, -, -, 1178739757, 11/12?1=1&2=2&3=3&4=?, 302, 17, ?, -1, -1

# Autres propriétés

- Autres propriétés
  - utilisation d' URL encoding
    - les caractères '%xx' des URL (codage hexadecimal) sont retraduits en caractères ASCII (*e.g.* '%3F' = '?')
  - lecture des formats de logs Apache les plus utilisés : 'common' et 'combined'

- voir readme.txt
- aide sur l'utilisation : option '-h'

# Performances

- 1,4 M de lignes traités en 3'51"
- Traitement en flux => pas d'explosion de la mémoire requise
- Filtrage efficace :
  - Si on s'intéresse qu'aux scripts, on retient environ 4% des données initiales

# Log Analyzer

- Outil graphique pour la visualisation, l'exploration de logs Apache
  - Facilite l'identification d'intrusions!

- Outil devant servir à l'évaluation de méthodes de détection d'intrusions à partir de logs

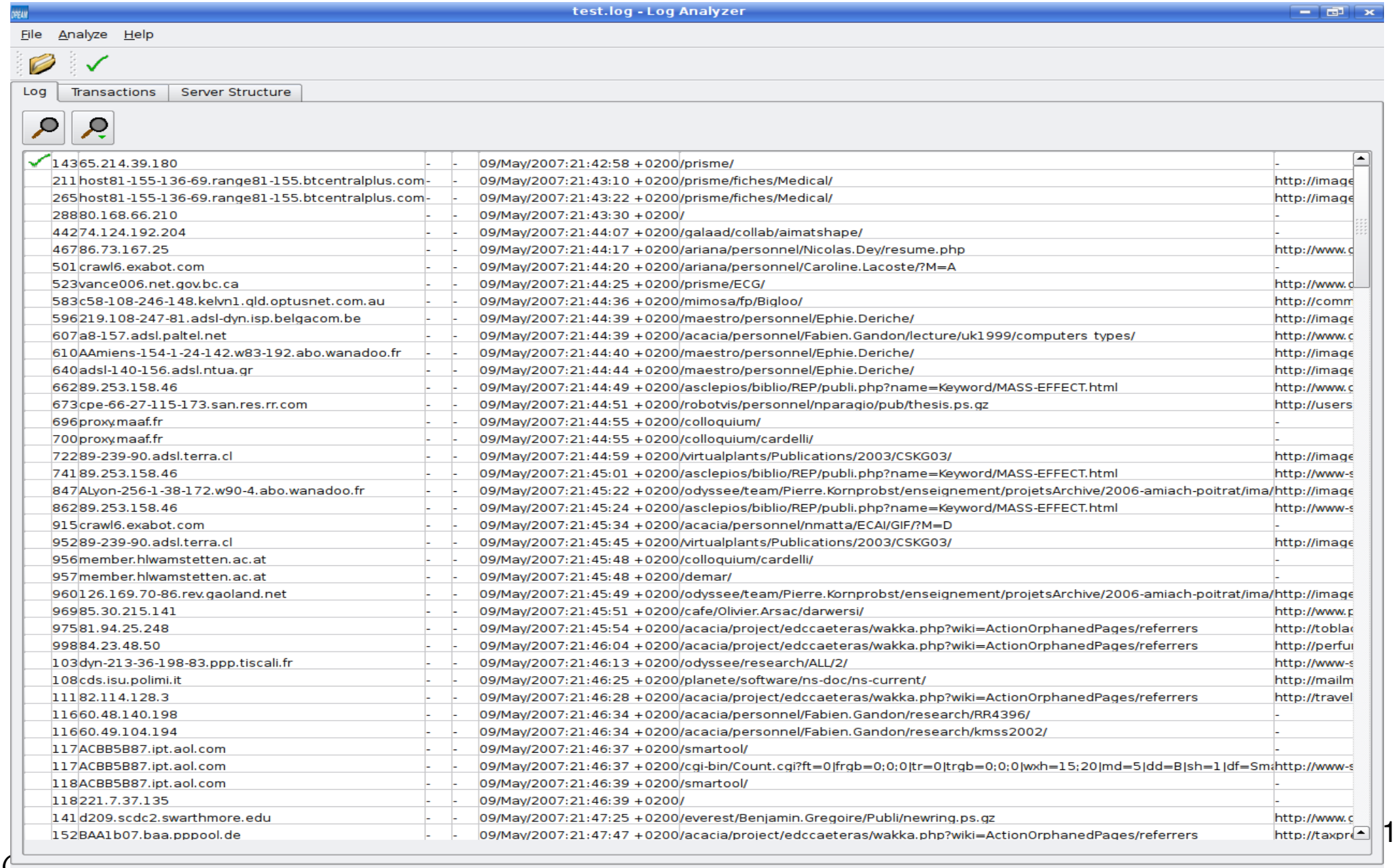
- Nom plus adéquate ??!!

# Log Analyzer

- Chargement de logs
  - L'intégralité des informations est conservée en mémoire : limite la capacité de traitement !
  - Utilisation nécessaire du filtrage *a priori*

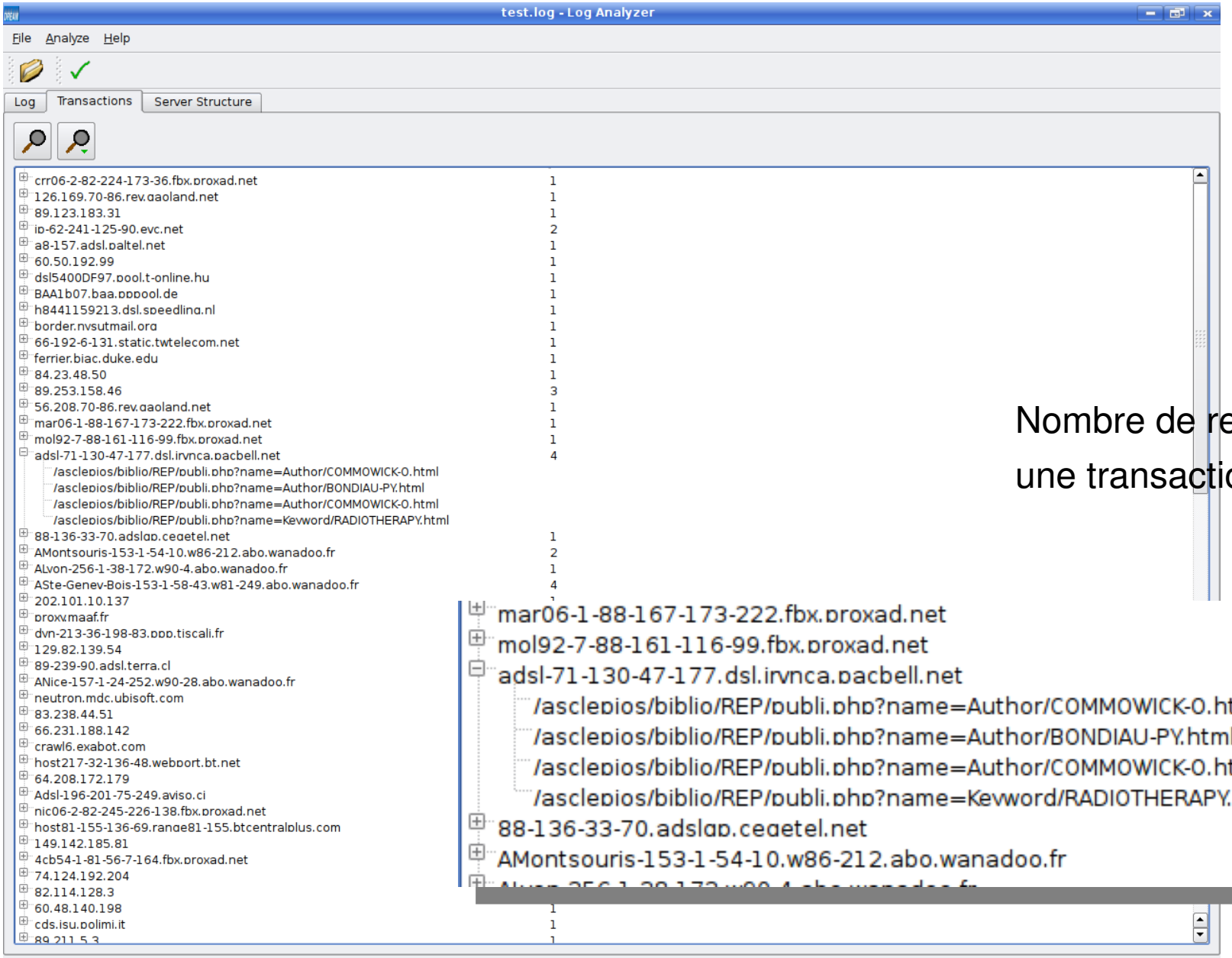
- Construction :
  - Transactions : regroupement des requêtes réalisées par un unique utilisateur (identifié par IP, sans limite de temps)
  - Image du serveur : reconstruction de la structure du serveur à partir des requêtes réalisées

# Log Analyzer : le log filtré



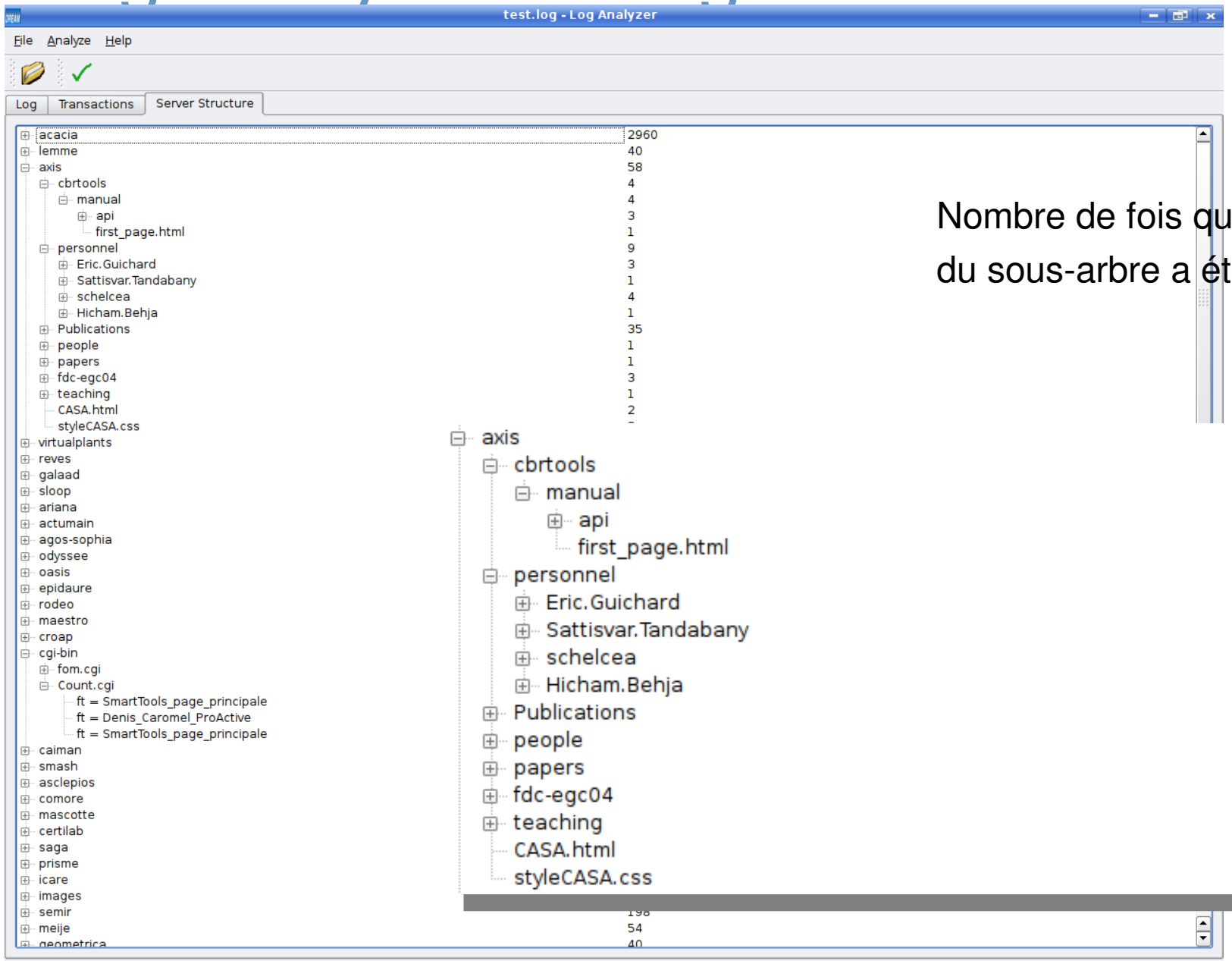

| | | | | | | |
|---|---|---|---|---|---|---|
| 143 | 65.214.39.180 | - | - | 09/May/2007:21:42:58 +0200 | /prisme/ | - |
| 211 | host81-155-136-69.range81-155.btcentralplus.com | - | - | 09/May/2007:21:43:10 +0200 | /prisme/fiches/Medical/ | http://image |
| 265 | host81-155-136-69.range81-155.btcentralplus.com | - | - | 09/May/2007:21:43:22 +0200 | /prisme/fiches/Medical/ | http://image |
| 288 | 80.168.66.210 | - | - | 09/May/2007:21:43:30 +0200 | / | - |
| 442 | 74.124.192.204 | - | - | 09/May/2007:21:44:07 +0200 | /galaad/collab/aimatshape/ | - |
| 467 | 86.73.167.25 | - | - | 09/May/2007:21:44:17 +0200 | /ariana/personnel/Nicolas.Dey/resume.php | http://www.c |
| 501 | crawl6.exabot.com | - | - | 09/May/2007:21:44:20 +0200 | /ariana/personnel/Caroline.Lacoste/?M=A | - |
| 523 | vance006.net.gov.bc.ca | - | - | 09/May/2007:21:44:25 +0200 | /prisme/ECG/ | http://www.c |
| 583 | c58-108-246-148.kelvn1.qld.optusnet.com.au | - | - | 09/May/2007:21:44:36 +0200 | /mimosa/fp/Bigloo/ | http://comm |
| 596 | 219.108-247-81.adsl-dyn.isp.belgacom.be | - | - | 09/May/2007:21:44:39 +0200 | /maestro/personnel/Ephie.Deriche/ | http://image |
| 607 | a8-157.adsl.paltel.net | - | - | 09/May/2007:21:44:39 +0200 | /acacia/personnel/Fabien.Gandon/lecture/uk1999/computers_types/ | http://www.c |
| 610 | AAmiens-154-1-24-142.w83-192.abo.wanadoo.fr | - | - | 09/May/2007:21:44:40 +0200 | /maestro/personnel/Ephie.Deriche/ | http://image |
| 640 | adsl-140-156.adsl.ntua.gr | - | - | 09/May/2007:21:44:44 +0200 | /maestro/personnel/Ephie.Deriche/ | http://image |
| 662 | 89.253.158.46 | - | - | 09/May/2007:21:44:49 +0200 | /asclepios/biblio/REP/publi.php?name=Keyword/MASS-EFFECT.html | http://www.c |
| 673 | cpe-66-27-115-173.san.res.rr.com | - | - | 09/May/2007:21:44:51 +0200 | /robotvis/personnel/nparagio/pub/thesis.ps.gz | http://users |
| 696 | proxy.maaf.fr | - | - | 09/May/2007:21:44:55 +0200 | /colloquium/ | - |
| 700 | proxy.maaf.fr | - | - | 09/May/2007:21:44:55 +0200 | /colloquium/cardelli/ | - |
| 722 | 89-239-90.adsl.terra.cl | - | - | 09/May/2007:21:44:59 +0200 | /virtualplants/Publications/2003/CSKG03/ | http://image |
| 741 | 89.253.158.46 | - | - | 09/May/2007:21:45:01 +0200 | /asclepios/biblio/REP/publi.php?name=Keyword/MASS-EFFECT.html | http://www-s |
| 847 | ALyon-256-1-38-172.w90-4.abo.wanadoo.fr | - | - | 09/May/2007:21:45:22 +0200 | /odyssee/team/Pierre.Kornprobst/enseignement/projetsArchive/2006-amiach-poitrat/ima/ | http://image |
| 862 | 89.253.158.46 | - | - | 09/May/2007:21:45:24 +0200 | /asclepios/biblio/REP/publi.php?name=Keyword/MASS-EFFECT.html | http://www-s |
| 915 | crawl6.exabot.com | - | - | 09/May/2007:21:45:34 +0200 | /acacia/personnel/nmatta/ECAI/GIF/?M=D | - |
| 952 | 89-239-90.adsl.terra.cl | - | - | 09/May/2007:21:45:45 +0200 | /virtualplants/Publications/2003/CSKG03/ | http://image |
| 956 | member.hlwamstetten.ac.at | - | - | 09/May/2007:21:45:48 +0200 | /colloquium/cardelli/ | - |
| 957 | member.hlwamstetten.ac.at | - | - | 09/May/2007:21:45:48 +0200 | /demar/ | - |
| 960 | 126.169.70-86.rev.gaoland.net | - | - | 09/May/2007:21:45:49 +0200 | /odyssee/team/Pierre.Kornprobst/enseignement/projetsArchive/2006-amiach-poitrat/ima/ | http://image |
| 969 | 85.30.215.141 | - | - | 09/May/2007:21:45:51 +0200 | /cafe/Olivier.Arsac/darwersi/ | http://www.p |
| 975 | 81.94.25.248 | - | - | 09/May/2007:21:45:54 +0200 | /acacia/project/edccaeteras/wakka.php?wiki=ActionOrphanedPages/referrers | http://tobla |
| 998 | 84.23.48.50 | - | - | 09/May/2007:21:46:04 +0200 | /acacia/project/edccaeteras/wakka.php?wiki=ActionOrphanedPages/referrers | http://perfu |
| 103 | dyn-213-36-198-83.ppp.tiscali.fr | - | - | 09/May/2007:21:46:13 +0200 | /odyssee/research/ALL/2/ | http://www-s |
| 108 | cds.isu.polimi.it | - | - | 09/May/2007:21:46:25 +0200 | /planete/software/ns-doc/ns-current/ | http://mailm |
| 111 | 82.114.128.3 | - | - | 09/May/2007:21:46:28 +0200 | /acacia/project/edccaeteras/wakka.php?wiki=ActionOrphanedPages/referrers | http://travel |
| 116 | 60.48.140.198 | - | - | 09/May/2007:21:46:34 +0200 | /acacia/personnel/Fabien.Gandon/research/RR4396/ | - |
| 116 | 60.49.104.194 | - | - | 09/May/2007:21:46:34 +0200 | /acacia/personnel/Fabien.Gandon/research/kmss2002/ | - |
| 117 | ACBB5B87.ipt.aol.com | - | - | 09/May/2007:21:46:37 +0200 | /smartool/ | - |
| 117 | ACBB5B87.ipt.aol.com | - | - | 09/May/2007:21:46:37 +0200 | /cgi-bin/Count.cgi?ft=0\|frgb=0;0;0\|tr=0\|trgb=0;0;0\|wxh=15;20\|md=5\|dd=B\|sh=1\|df=Sm | http://www-s |
| 118 | ACBB5B87.ipt.aol.com | - | - | 09/May/2007:21:46:39 +0200 | /smartool/ | - |
| 118 | 221.7.37.135 | - | - | 09/May/2007:21:46:39 +0200 | / | - |
| 141 | d209.scdc2.swarthmore.edu | - | - | 09/May/2007:21:47:25 +0200 | /everest/Benjamin.Gregoire/Publi/newring.ps.gz | http://www.c |
| 152 | BAA1b07.baa.pppool.de | - | - | 09/May/2007:21:47:47 +0200 | /acacia/project/edccaeteras/wakka.php?wiki=ActionOrphanedPages/referrers | http://taxpr |

# Log Analyzer : les transactions




Nombre de requêtes dans une transaction

# Log Analyzer : image du serveur




Nombre de fois qu'une page du sous-arbre a été accédé

# Log Analyzer : exemple d'utilisation

- Recherche d'intrusions
  - A partir de l'image du serveur, on peut identifier s'il y a eu des requêtes à des scripts « sensibles » (*e.g.* `passwd.cgi`)
  - Si elle existe, on peut s'intéresser à les repérer dans le log
  - Et également identifier, avec les transactions, qui les a faite et s'il s'agit bien d'une intrusion!

- Très efficace pour identifier `salmacis` : intrusion pendant 3 min, identifiée dans 3 jours de logs.

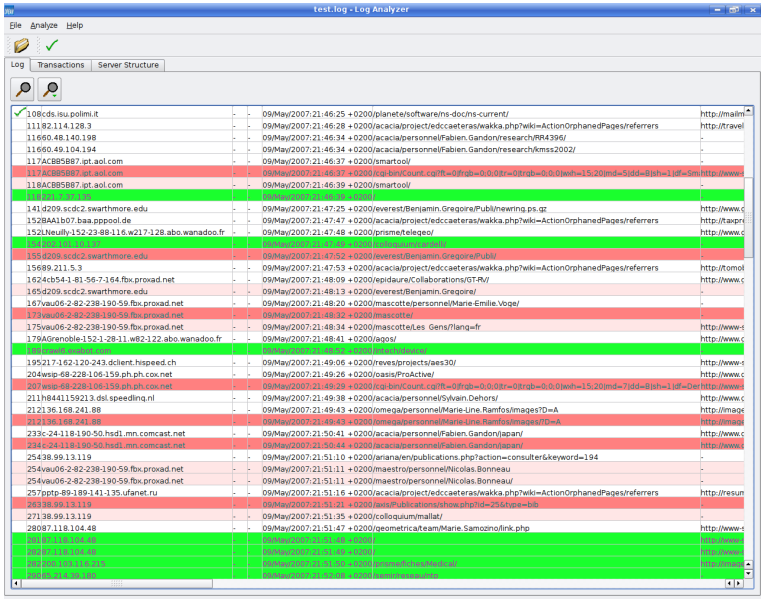
# Log Analyzer : analyse de logs

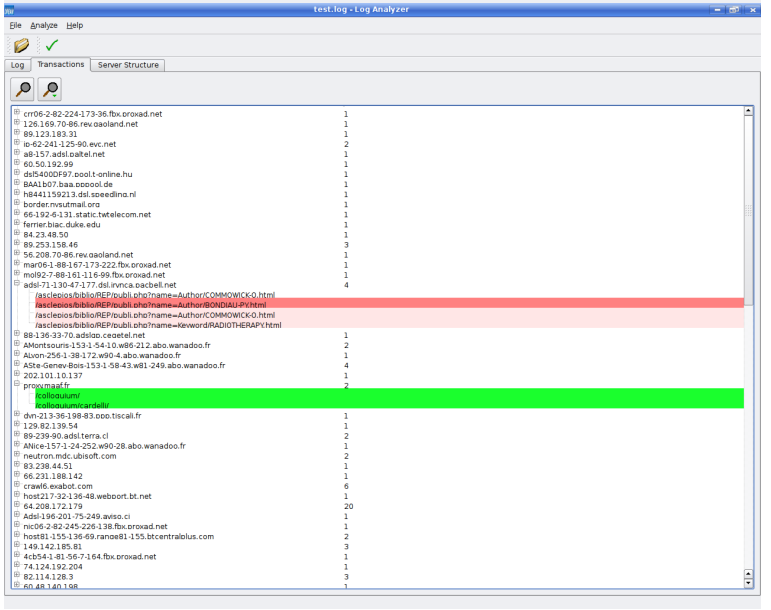
- Permet l'implémentation de méthodes d'analyse de lignes de log ou de transactions
  - Mise à disposition de classes abstraites dans le code
- L'outil facilite la visualisation des résultats et l'identification des qualités et défauts de chaque modèle et méthode utilisés
- Aspect data stream :
  - Données traitées dans l'ordre
  - Mais, pas de contrainte de temps de calcul
  - Pourrait être simulé

# Log Analyzer : résultat d'analyse

- Visualisation :
  - Rouge = intrusion
  - Vert = normal
- Exemple : utilisation de distribution de caractères
  - Adaptation du modèle de distribution lorsqu'il détecte trop d'intrusions

Exemple illustratif pour la visualisation, mais mauvais pour la détection d'intrusion



# Log Analyzer : version Beta

- La version développée est fonctionnelle
- Utilise les librairies Qt (portabilité sous Windows)
- Une version améliorée est en cours
  - Éclaircissement du code (!)
  - Amélioration de la gestion de la mémoire
- Autres fonctionnalités éventuelles ... sur propositions
  - Extension pour l'annotation manuelle des lignes de log (?)

# Robot web

- Objectif:
  - Détection d'intrusions par liste blanche (maintenue à jour automatiquement)
  - Le robot recueille l'ensemble des pages valides du site

- Simplification de l'outil `htdig`
  - Htdig : outil open source pour la création de moteur de recherche sur serveurs web.
  - Permet de parcourir un site web : html, scripts serveur (*e.g.* php, cgi), pdf, swf, ...

- Récupération du robot, et suppression des parties non-utiles